TECHNICAL REPORT

DSPL-85-1

IMPROVED OBJECTIVE MEASUREMENTS FOR SPEECH QUALITY TESTING

THOMAS P. BARNWELL III

MARK A. CLEMENTS

SCHULYER R. QUACKENBUSH

ERIC P. FARGES

JANUARY 1985

GEORGIA INSTITUTE OF TECHNOLOGY

A UNIT OF THE UNIVERSITY SYSTEM OF GEORGIA
SCHOOL OF ELECTRICAL ENGINEERING
ATLANTA, GEORGIA 30332

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: DCA100-83-C-0027
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): Improved Objective Measures for Speech Quality Testing
5. TYPE OF REPORT & PERIOD COVERED
6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): Thomas P. Barnwell III, Mark A. Clements, Schuyler R. Quackenbush, and Eric P. Farges
8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS:
Defense Communications Engineering Center
1860 Wiehle Av.
Reston, Virgina 22090
12. REPORT DATE
13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
15. SECURITY CLASS. (of this report)
15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Unlimited, Open Publication

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
Same

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Speech, Speech Quality, Quality, Objective Testing, Subjective Testing, Voice Communication, Speech Communications, Coding, Speech Coding.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
This report presents the results of a large study of new objective measures for speech quality. The study defined a figure-of-merit for a particular objective quality measure using correlation analysis between a data base of subjective speech quality measures and a data base of objective speech quality measures. Both data bases were derived from approximately twenty hours of coded and distorted speech. The subjective test used was the Diagnostic Acceptability

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Measure (DAM) developed at the Dynastat Corporation.

The study consentrated on three classes of objective measures: those based on a model of the ear, those based on parametric subjective quality estimates, and those based on preclassified distortion sets. All three classes of objective measures performed well, but the best results were obtained for the class of parametric measures.

IMPROVED OBJECTIVE MEASURES FOR SPEECH QUALITY TESTING

By

Thomas P. Barnwell, III
Mark A. Clements
Schuyler R. Quackenbush
Eric P. Farges

Digital Signal Processing Laboratory
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

FINAL REPORT

Prepared for

Defense Communications Agency
Defense Communications Engineering Center
1860 Wiehle Avenue
Reston, VA 22090

September 1984

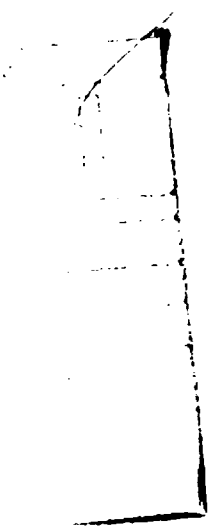


TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES CONTINUED

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Task History

The research effort reported here was performed in the Digital Signal Processing Laboratory of the School of Electrical Engineering at the Georgia Institute of Technology. In this effort, the Georgia Institute of Technology was the prime contractor and the Dynastat Corporation of Austin, Texas operated as a subcontractor. The monitoring officer at the Defense Communications Engineering Center was Mr. Kenneth Fischer.

This task, which sought to develop new compactly computable objective measures for the prediction of subjective quality assessments of speech coding systems, followed previous work by both Georgia Tech [1.1-1.13] and the Dynastat Corp. [1.5] [1.14] [1.15] in relate areas. In this study, all of the research work was performed at Georgia Tech, while the Dynastat Corporation's sole function was to perform the required subjective quality evaluations.

1.2 Technical Background

In recent years, considerable effort has been devoted to the development of efficient digital speech coding algorithms for the transmission and storage of speech signals. These algorithms represent a wide range of approaches to the speech coding problem, and a correspondingly wide range of data rates, computational intensities, and perceived distortion characteristics. At the high data rates, such simple systems as mu-law and A-law PCM coders operate with toll quality at around 64K bps. At intermediate rates (32K bps-9.6K bps) such systems as DM [1.16], ADM [1.17][1.18], DPCM [1.19], ADPCM [1.20], APC [1.21], SBC [1.22], and ATC [1.23][1.24] are currently being used and proposed. In addition 'gapped analysis' [1.20][1.25] or 'harmonic scaling' [1.26] is also

effective in reducing bit rates in this range. At the lower data rates (2.4K bps-200 bps), fixed rate pitch excited LPC [1.27-1.29] and channel [1.30-1.32] vocoders are being used, and variable rate [1.33][1.34], vector quantized [1.35][1.36], and recognition/synthesis [1.37][1.38] systems are being proposed. In addition, considerable progress is now being made in the 0.6-2.4K bps range by such techniques as noise feedback [1.39] and run-length-coding [1.40] in APC and parametric excitation representations in residual excited vocoders [1.41][1.42].

The problem of rating and comparing these systems from the standpoint of user acceptance is a difficult one, since the candidate systems are usually highly intelligible. Hence, context free intelligibility tests such as the DRT [1.47] and the MRT [1.48] may not suffice to resolve small differences in acceptability. User preference tests, such as the PARM [1.15], the QART [1.15], and the more modern DAM [1.16] can be effective in assessing quality, but they all suffer from the inherent drawbacks of subjective tests. These include both the great care which must be exercised to obtain repeatable subjective results and the corresponding expense and lack of flexibility associated with such testing.

Objective acceptability measures, on the other hand, do not suffer from many of the problems of subjective tests [1.1-1.13]. On the whole, they are easy to administer and many have proved to be very reliable [1.15]. Likewise, many objective measures can be implemented in real-time or near-real-time, which vastly extends their flexibility. Also, objective measures may often be used directly in the design of speech coding systems in ways which are not possible with subjective measures.

The problem is that it would be naive to believe that any simple, compactly computable objective measure could be designed which would always correlate well with subjective quality results across a large ensemble

of coding and other distortions. Despite our poor understanding of the speech perception process at present, we can assuredly state that the human listener is an active perceiver who uses his immense knowledge of the language, the talker, and the semantic and syntactic context to 'fill in the gaps' in the perceived speech. Hence, it is clear that no objective measure which does not use semantic, syntactic, and talker related information can ever be expected to perform well across all possible speech distortions, and such measures are clearly not possible with our current knowledge. On the other hand, it is fair to say that with the possible exception of very low bit rate recognition/synthesis systems, the distortions found in speech coding systems are not synchronized with the semantic, syntactic, or talker related features of the speech signal.

The challenge in the design of compactly computable objective measures is hence to realize maximum utility from a set of intrinsically imperfect procedures. Until recently, the relative performance of different objective measures in terms of their ability to predict subjective quality results has not been well understood. However, in a recent study funded by the Defense Communications Agency (DA100-78-C-003) [1.5] and later by the National Science Foundation (ECS-801-6712) the relative performances of many objective speech quality measures have been addressed in detail [1.1-1.13]. In many ways, the research which is being reported in this document can be considered to be a continuation of these studies.

1.3 The Technical Approach

In the earlier research, the emphasis was on comparing and quantifying the performance of a large number of parametric variations of simple objective measures. The basic methodology employed in both the earlier research and in this research, which is based on correlation analyses between objective and

subjective speech quality measures applied across a large ensemble of coded and distorted speech, is described in detail in Chapter 2 of this report. At onset of this research, about 2000 objective measures had been studied using about 140,000 individual correlation analyses.

The experimental and research environment developed in the previous research efforts offers a unique opportunity for the design, implementation, and evaluation of new, more complex objective speech quality measures. On the one hand, the body of the research performed over the last five years has provided a good understanding of the relative performance of a large number of individual objective measures. On the other hand, the experimental environment itself both offers an efficient method for testing objective measures and also represents an outstanding resource for the design of new objective measures. In this context, the goal of this research was to use the existing resources to maximum advantage in developing and evaluating a new set of objective measures for the efficient prediction of the user acceptance of speech coding systems.

Two particular application areas for objective quality measures are particularly appropriate to the concerns of the Defense Communications Agency. The first is the area of designing devices for field testing the performance of digital coding systems which are either being installed or which may have been degraded by system failures. The second is the area of developing techniques to be used in conjunction with subjective quality measures for improving the resolving power or reducing the cost of system acceptability assessment. This research explicitly addressed both of these areas.

The constraints imposed by the two applications areas are quite different. Algorithms to be used by quality assessment devices in the field must be compactly computable to allow for their implementation on modern signal processing hardware. Likewise, they should be extremely sensitive to any system degradation, and should indicate with high resolution whether the system

is 'working correctly'. In addition, if the system quality is degraded, such algorithms should give good estimates of the loss of quality due to the degradation. The general applicability of these devices to a very large class of coding systems is of secondary concern in this environment, since the ensemble of coding systems is limited. The key research question in this area, therefore, is given computational constraints, how large a class of distorting systems can be effectively addressed by composite objective measures.

On the other hand, algorithms to be used primarily for quality assessments must conform to a different set of constraints. First, of course, since they may be performed in non-real-time, they may be moderately computationally intense (as compared to the highly computationally intense iterative measures employed in digital coder design). Likewise, they must address a far broader range of distortions if they are to be effective. In this regard, it may be possible to develop objective measures tuned to some general distortion characteristics (e.g. waveform coders, pitch excited vocoders, or frequency domain coders), but any such dynamic variation in the application of the objective measure algorithm must also be driven objectively. To design such measures effectively, it is important to configure the algorithms in a perceptually relevant way. Stated another way, if a broad class of distortions are to be included, objective measures should be designed to estimate quantities which are directly related to the quality degradations perceived by humans.

The design of objective speech quality measures for these two applications areas were addressed in the context of a three part study. Although in some sense all three parts address both application areas, the first two parts were particularly intended to address issues germane to the general quality assessment problem, while the third part addressed the field quality testing

1.4 Objective Measures Based On Signal Processing Models For The Inner Ear

The first part of the research dealt specifically with designing new objective speech quality measures based on signal processing models for the inner ear. A detailed description of this research and its results is given in Chapter 4.

Briefly, the question of designing and assessing objective measures based on aural models was addressed in a three phase study. In the first phase, models related to those previously proposed along with possible augmentations were studied, and a set of parameterized objective measures were developed. In the second phase, the control parameter space was studied using correlation analysis techniques described in Chapter 2. In the final phase, the optimized objective measures from phase two were combined with other objective measures to form improved composite measures.

For the most part, the objective measures studied here can be considered to be parameterized, frequency-variant spectral distance measures. In the original research [1.5], the best of this class of measure was found to have a correlation coefficient of .60 across all distortions for frequency-invariant spectral distance measures, and a correlation coefficient of .71 for frequency-variant spectral distance measures. The new measures designed in this research were able to achieve a correlation coefficient of .78 across the same distortion ensemble. This can be considered to be a good, although no spectacular, improvement for this class of measure. The best results were obtained for measures designed using the principals first suggested by Dennis Klatt [1.49]. Based on these and other related results, it is a reasonable conjecture that the level of performance achieved here is near the maximum which can be expected from simple, fully parameterized spectral distance measures.

1.5 Parametric Objective Quality Measures

Two of the attractive features of the DAM [1.14] are that its parametric subjective quality estimates serve to give insight into the perceived nature as well as the perceived level of the distortion and the regression model which relates the parametric subjective qualities to the estimated system acceptability gives insight on the relative importance of different parametric qualities. If an objective measure is to succeed over a large class of distorting systems, then it must somehow incorporate information related to the perceived nature of the distortion.

Part two of this study was aimed at designing a better objective quality measure based on individual parametric objective measures. A detailed description of this research is given in Chapter 5. In the first phase of this study, multi-dimensional scaling was used to characterize the relationship between the objective measures previously designed, the isometric subjective speech quality measures, and the parametric subjective speech quality measures. This initial analysis proved to be the key to designing better objective measures in that it characterized the problem in such a way that the design issues became obvious. In the second phase, a regression analysis was performed which showed exactly which parametric measures are most important in predicting system acceptability. As a result of this regression analysis, a subset of parametric subjective measures was identified for further study. In the ensuing phases, a specific objective measure was designed to predict each of the parametric subjective measures in the subset. This design was done interactively using statistical analysis techniques on the speech quality data bases.

On the whole, the results of this part of the research were very good. In particular, it was possible to identify exactly where the previously proposed objective measure were breaking down, and further it was possible to see

exactly what had to be done to correct the problem. What had to be done, in this context, was to design particular new objective measures which predicted particular parametric speech quality measures. The result of this effort was a number of new parametric objective measures which did an exceptional job at predicting many of the important parametric subjective measures.

In effect, what has been designed and studied here is an objective version of the DAM. The test will provide an overall acceptability estimate and set of parametric quality estimates for individual perceived qualities. It would be naive, of course, to expect such a measure to perform comparably with the DAM itself. However, such a test along with a complete statistical analysis of its projected performance, should prove very valuable in both screening of systems before the application of subjective quality tests and in providing analytically tractable information on the nature of the distortion for use in the coder design problem.

It would be misleading to imply that this study was completely successful. In particular, the performance of the new parametric objective measures was varied, and whereas some performed extremely well, others were not as successful. Nevertheless, it is fair to say that these results represent a major improvement in our understanding and our ability to implement objective speech quality measures.

1.6 Classified Objective Measures

The third part of the research was a systematic study of classified objective measures as applied to distortion subsets. A classified objective measure is one which performs differently based on 'classification information' which is available. This information may be an external input to the measure (such as an operator supplied classification) or it may be an internally supplied parameter (such as an objective classification of sound segments into

approximate linguistic categories). The details of this research are found in Chapter 6.

The research on classified objective measures really had two goals. The first goal was to investigate the use of classified measures for very narrow classes of measures. The purpose of this part of the study was to design measures appropriate for field testing communications systems where the class of system in use was known. The second goal was to design new, broad based classified measures for a large ensemble of distortions. The basic approach used in this part of the research was to use statistical techniques to identify distortion subsets for which the subjective measures could be predicted well by the objective measures under study.

It is fair to say that the research on the classified objective measures was the least successful of the three approaches. It is true that the work clearly illustrated the viability of using narrowly classified objective measures for field testing applications. It is also true that it was clearly illustrated that the distorted data base could be partitioned so that high quality classified objective measures could be designed for use with a large distortion ensemble. The problem was that the members of the required distortion subsets appeared to be so dissimilar in both their perceptual characteristics and their signal characteristics that we were unable to adequately specify either objective or subjective rules for classifying the distortion. This does not really prove that this approach is without merit. It means, rather, that at this time we have not been able to discover distortion classification techniques which work well enough to prove out the approach.

The Distortion Ensemble Augmentation

The final task which was addressed as part of this research contract was the augmentation of the existing distortion ensemble from 264 distortions to

318 distortions. Fundamentally , two classes of distortions were included in these new distortions. The first were a set of speech coding techniques which had been developed and become common since the original data bases were developed in 1978. These new coding distortions included subband coders, adaptive transform coders, ADPCM with noise feedback, multi-pulse LPC, and channel vocoders. The second were a new set of 'banded pole distortion' controlled distortions [1.5]. The purpose of these new controlled distortions was to increase the overall spread of subjective responses, which had been inadequate in the first study. The new coding and controlled distortions are described in detail in Chapter 3.

The basic design criterion for all of the distortions was to have each range from 'barely perceivable' to 'moderately distorted'. All of the new distortions met this criterion with the possible exception of the channel vocoder, for which the spread in subjective responses was slightly less than desired.

REFERENCES

[1.1] T.P. Barnwell and W.D. Voiers, 'An Analysis of Objective Measures for User Acceptance of Voice Communications Systems,' Final Report to the Defense Communications Agency, DCA100-78-C-0003, September 1979.

[1.2] T.P. Barnwell, III, A.M. Bush, R.M. Mersereau, and R.W. Schafer, 'Speech Quality Measurement,' Final Report DCA/DCEC F30602-77-C-0118, June 1977.

[1.3] T.P. Barnwell, III, R.W. Schafer, and A.M. Bush, 'Tandem Interconnections of LPC and CVSD Digital Speech Coders,' Final Report, DCA 100-76-6-0073, 15 November 1977.

[1.4] T.P. Barnwell, III and A.M. Bush, 'Statistical Correlation Between Objective and Subjective Measures for Speech Quality,' 1978 International Conference on Acoustics, Speech, and Signal Processing, April 1978.

[1.5] T.P. Barnwell and W.D. Voiers, 'Objective Fidelity Measures for Speech Coding Systems,' presented at the meeting of the Acoustical Society of America, Honolulu, December 1978.

[1.6] T.P. Barnwell, 'Objective Fidelity Measures for Speech Coding Systems,' Acoustical Society of America, Vol. 65, No. 6, December 1979.

[1.7] T.P. Barnwell, 'Correlation Analysis of Subjective and Objective Measures for Speech Quality,' 1980 International Conference on Acoustics, Speech, and Signal Processing, Denver, Colorado, April 1980.

[1.8] T.P. Barnwell, 'A Comparison of Parametrically Different Objective Speech Quality Measures Using Analysis with Subjective Quality Results,' 1980 International Conference on Acoustics, Speech, and Signal Processing, Denver, Colorado, April 1980.

[1.9] T.P. Barnwell and P. Breitkopf, 'Segmental Preclassification for Improved Objective Speech Quality Measures,' Proc. of ICASSP '81, March 1981.

[1.10] T.P. Barnwell, III, 'On the Standardization of Objective Measures for Speech Quality Testing,' Proceedings of 1982 NBS Workshop on Standards for Speech Recognition and Synthesis, Washington, DC, March 1982.

[1.11] T.P. Barnwell, III, and S.R. Quackenbush, 'An Analysis of Objectively Computable Measures for Speech Quality Testing,' Proc. of ICASSP '82, May 1982.

[1.12] S.R. Quackenbush and T.P. Barnwell, III, 'An Approach to Formulating Objective Speech Quality Measures,' Proc. 15th Southeastern Symposium on System Theory, Huntsville, Alabama, March 28-29, 1983.

[1.13] S.R. Quackenbush and T.P. Barnwell, III, 'The Estimation and Evaluation

of Pointwise Nonlinearities for Improving the Performance of Objective Speech Quality Measures,' Proc. ICASSP '83, Boston, Mass., April 1983.

[1.14] W.D. Voiers, 'Diagnostic Acceptability Measure for Speech,' 1977 International Conference on Acoustics, Speech, and Signal Processing, Hartford, CN, May, 1977.

[1.15] W.D. Voiers et al., 'Methods of Predicting User Acceptance of Voice Communications Systems,' Final Report, DCA 100-74-C-0056, DCA, DCEC, Reston, VA, July 1976.

[1.16] N. S. Jayant, 'Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers,' Proceedings of IEEE , May 1974.

[1.17] N. S. Jayant, P. Cummiskey, and J. L. Flanagan, 'Design and Implementation of an Adaptive Delta Modulator,' Proc. of IEEE Int. Conf. Speech Communications , Boston, MA, April 1972.

[1.18] N. S. Jayant, 'Adaptive Delta Modulator with a One-Bit Memory,' Bell System Tech. Journal , vol. 49, March 1970.

[1.19] M. D. Paeb and T. H. Glisson, 'Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM,' IEEE Trans. on Comm., April 1972.

[1.20] T. P. Barnwell, III, A. M. Bush, J. B. O'Neal, and P. W. Stroh, 'Adaptive Differential PCM Speech Transmission,' Final Report to the Defense Communications Agency, RADC-TR-74-177, July 1974.

[1.21] B. S. Atal and M. R. Schroeder, 'Adaptive Predictive Coding of Speech Signals,' Bell System. Tech. Journal , October 1970.

[1.22] R. E. Crochiere, S. A. Webber, and J. L. Flanagan, 'Digital Coding of Speech in Sub-bands,' Proc. 1976 IEEE Int. Conf. on ASSP, pp. 233-236, March 1976.

[1.23] R. Zelinski and P. Noll, 'Approaches to Adaptive Transform Coding of Speech at Low Rates,' IEEE Trans. on Acoustics, Speech, and Signal Processing , vol. ASSP-27, no. 1, Feb. 1979.

[1.24] J. M. Tribolet and R. E. Crochiere, 'Frequency Domain Coding of Speech,' IEEE Trans. on ASSP , vol. ASSP-27, no. 5, October 1979.

[1.25] T. P. Barnwell, III and A. M. Bush, 'Gapped ADPCM for Speech Digitization,' Proc. of NEC , October 1974.

[1.26] D. Malah, 'Time Domain Algorithm for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals,' IEEE Trans. on ASSP , vol. ASSP-27, April 1979.

[1.27] B. S. Atal and S. L. Hanaver, 'Speech Analysis and Synthesis by Linear Prediction of the Speech Waveform,' Journal of Acoustical Soc. of America , vol. 50, 1971.

[1.28] F. Itakura and S. Saito, 'Analysis Synthesis Telephony Based on the Maximum Likelihood Method,' Proc. Sixth Int. Congr. Acoust. , 1968.

[1.29] J. Makhoul, 'Linear Prediction: A Tutorial Review,' Proc. IEEE , vol. 63, 1975.

[1.30] H. Dudley, 'Remaking Speech,' J. Acoust. Soc. Am. , vol. 11, 1939a.

[1.31] B. Gold and C. M. Rader, 'The Channel Vocoder,' IEEE Trans. on Audio and Electroacoustics , vol. AU-15, no. 4, pp. 148-160, Dec. 1967.

[1.32] J. N. Holmes, 'Dynamic Encoding as Applied to a Channel Vocoder,' IEEE Trans. Comm. Syst. , vol. 11, 1963.

[1.33] D. T. Magill, 'Adaptive Speech Compression for Pocket Communications Systems,' Telecommun. Conf. Rec. , IEEE Publ. 73 CHO 805-2, 29D 1-5, 1973.

[1.34] P. E. Papamichalis and T. P. Barnwell, III, 'A Dynamic Programming Approach to Variable Rate Speech Compression,' Proc. 1980 Int. Conf. ASSP , Denver, CO, April 1980.

[1.35] A. Buzo, A. H. Gray, R. M. Gray, and J. D. Markel, 'Speech Coding Based Upon Vector Quantization,' IEEE Trans. on ASSP, vol. ASSP-28, no. 5, pp. 562-547, October 1980.

[1.36] D. Wong, B. H. Juang, and A. H. Gray, 'Recent Developments in Vector Quantization for Speech Processing,' Proc. 1981 Int. Conf. on ASSP , pp. 1-4, Atlanta, GA, April 1981.

[1.37] B. T. Oshika, 'FACP Speech Recognition/Transmission Systems,' Final Technical Report, RADC-TR-78-193, System Development Corporation, August 1978.

[1.38] R. Schwartz, J. Klovstad, J. Makhoul, and J. Sorensen, 'A Preliminary Design of a Phonetic Vocoder Based on a Diphone Model,' Proc. 1980 Int. Conf. on ASSP , pp. 32-35, Denver, CO, April 1980.

[1.39] B. S. Atal and M. R. Schroeder, 'Improved Quantizer for Adaptive Coding of Speech Signals at Low Rates,' Proc. 1980 Int. Conf. on ASSP , pp. 535-538, Denver, CO, April 1980.

[1.40] M. R. Schroeder, 'Predictive Coding of Speech Signals and Subjective Error Criteria,' Trans. 1978 Int. Conf. on ASSP , pp. 573-576, 1978.

[1.41] B. S. Atal, 'A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates,' Proc. 1982 Int. Conf. on ASSP , pp. 614-617, Paris, France, May 1982.

[1.42] L. B. Almeida and J. M. Tribolet, 'A Spectral Model for Nonstationary Voiced Speech,' Proc. of 1982 Int. Conf. on ASSP , pp. 1303-1306, Paris, France, May 1982.

[1.43] B. S. Atal, 'Efficient Coding of LPC Parameters by Temporal Decomposition,' Proc. ICASSP 1983, pp 81-84.

[1.44] M. Berouti, H. Garten, P. Kabal, and P. Mermelstein, 'Efficient

Computation and Ending of the Multipulse Excitation for LPC,' Proc. ICASSP 1984, pp 10.1-10.4.

[1.45] G. A. Senenseib, A. J. Milbourn, A. H. Lloyd, and I. M. Warrington, 'A Non-Iterative Algorithm for Obtaining Multipulse Excitation for Linear Predictive Speech Coders,' Proc. Icassp 1984, pp 10.5-10.8.

[1.46] I. M. Trancoso, R. Garcia-Gomez, and J. M. Tribolet, 'A Study of Short Time Phase and Multipulse LPC,' Proc. ICASSP 1984, pp 10.9-10.12.

[1.47] W. D. Voiers, 'Research on Diagnostic DRT Evaluation of Speech Intelligibility,' Final Report AFSC No. F19628-70-C-0182, 1973.

[1.48] J. D. Griffiths, 'Rhyming Minimal Contrasts: A Simplified Diagnostic Articulation Test,' J. Acoust. Soc. Am. , vol. 42, no. 1, pp. 236-241, 1967.

[1.49] D.H. Klatt, 'Prediction of perceived phonetic distance from critical-band spectra: a first step,' Proceedings of International Conference on Acoustics, Speech and Signal Processing, 1982, Paris, pp. 1278-1281.

CHAPTER 2

THE TESTING OF OBJECTIVE MEASURES

2.1 Background

As was noted in the introduction, this research project is essentially a continuation of a research project funded by the Defense Communications Agency in 1978 entitled An Analysis of Objective Measures for User Acceptance of Voice Communications Systems [2.1]. The goal of the original work was to study the viability of using relatively simple, objectively computable measures for estimating the results of subjective speech quality tests. As part of the original research, a statistical technique for measuring the expected performance of objective speech quality measures was designed, implemented, and tested [2.1].

Much of the effort in the original research program was directed towards the goal of quantitatively evaluating the performance of many of the (relatively) simple objective quality measures which had been previously proposed and used in speech processing. The original study involved over 40,000 correlation analyses based on over 2000 separate objective speech quality measures. Most of these objective measures were parametric variations of compactly computable fidelity measures. The major accomplishment of this early work was that it gave for the first time a degree of quantitative insight into the way in which many objective measures performed relative to one another as well as to subjective quality estimates. This study showed, for example, that the relatively simple log area ratio measure performed as well as the more complex log spectral distance measures [2.1]. Likewise, the short-time frequency-variant SNR was found to be an outstanding measure for wave-form coders [2.1]. In addition, the effects of frequency variant [2.2][2.3] and time variant [2.4] objective measures were investigated in some detail. All of these

results served to provide much-needed insight into the fundamental nature of perception of speech distortion and the associated foundations of speech coder acceptability.

In another sense, however, the first study generated more questions than it answered. A basic feature of the approach used in both the current and original research programs is that the experimental procedure requires immense amounts of data reduction and data storage. This is a result of the very large size of the data bases involved (about 6×10^9 bytes of data storage) as well as the very large number of objective measures which can be studied in a single experiment. Stated simply, although it takes a great deal of effort to generate a single result, it takes little additional effort to generate many results. Hence, the experimenter is faced with the choice of either an intrinsically slow iterative design procedure or an immense data reduction task between experiments. As a result, the earlier research program was able to perform an extensive study of the class of simple objective speech quality measures, but it was only able to perform a limited study of the more complex and specialized measures. In particular, it performed an initial study of composite objective measures, which are single objective measures formed as combinations of several other objective measures, and parametric objective measures, which seek to estimate the parametric subjective speech qualities [2.1].

An important result of the original research program was that most of the simple objective measures currently in use, along with their parametric variations, do not perform very well when applied to a large class of dissimilar distorting systems. In particular, the highest correlation coefficient derived for a single, frequency-invariant objective measure applied across all distortions was in the range of .60 to .65 [2.1][2.2][2.5]. This

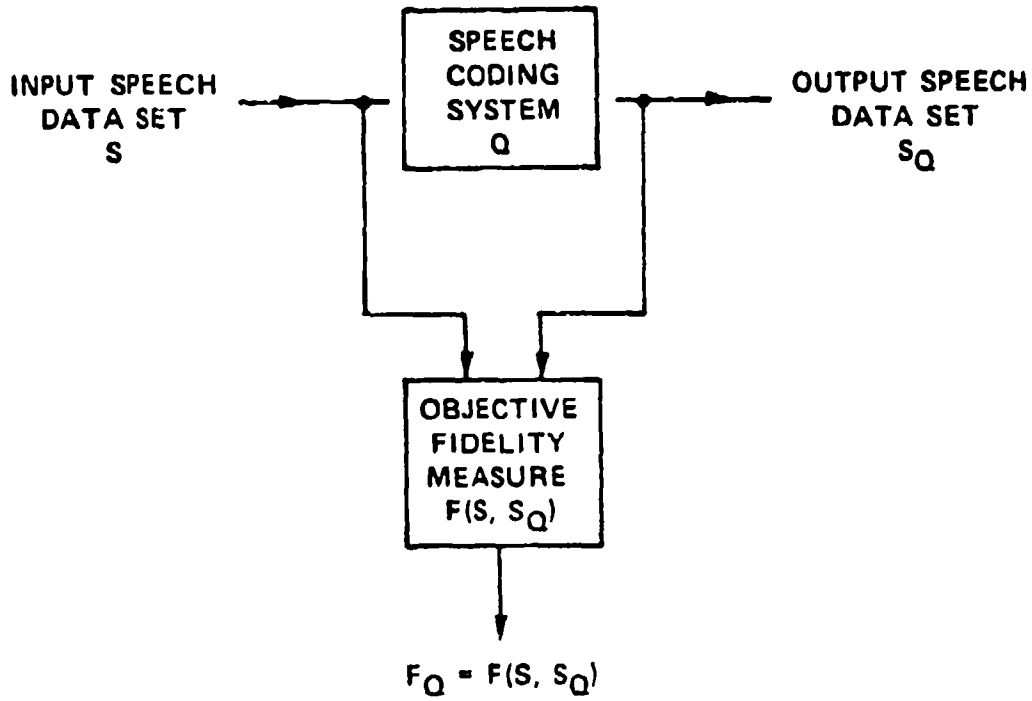
level of performance is not good enough to be of great utility for either quality assessment or coder design. However, a few initial experiments were performed on composite objective speech quality measures, which were formed as weighted sums of sets of dissimilar simple objective measures. Despite the fact that these early experiments used a broad statistical approach, which incorporated no special insight in regard to either the nature of the data bases or the nature of speech perception, the results were very promising. In particular, one composite measure was tested which attained a correlation score of .88 across the entire distortion ensemble. Because of the nature of the analysis procedures, however, it was not possible to interpret this result adequately in a broad sense. For example, the measure's robustness, as well as to what extent this measure's performance was due to the statistical properties of the data bases rather than fundamental properties of speech perception, are not clear.

In short, two basic points emerged from the results of the original research program. First, it seemed clear that new objective measures could be designed whose performance substantially exceeded the performance of the objective measures currently in use. Second, it also seemed clear that considerable additional work would be required in order to design these new measures. Due to the large size of the data bases involved and due to the computational intensity of the statistical estimation tasks, the original research had only begun the task of effectively using the data bases to design new objective speech quality measures. What was required was more in-depth look at the available data.

2.2 The Basic Testing Procedures

The objective speech quality measures of interest in this study can all be defined in terms of the model of Figure 2.2-1. In general, these objective measures are computed from an input or undistorted speech data set, S, and an

OBJECTIVE FIDELITY MEASURES



CONDITIONS FOR A MEASURE TO BE A METRIC

1. $F(S, S_Q) = F(S_Q, S)$
2. $F(S, S_Q) = 0$ if $S = S_Q$
   $F(S, S_Q) \geq 0$ if $S \neq S_Q$
3. $F(S, S_Q) \leq F(S, S_Y) + F(S_Y, S_Q)$

Figure 2.2-1. System for Computing Objective Quality Measures.

output or distorted speech data set, S_Q. The output speech data set is formed by passing the input speech data set through the speech communications system under test. It should be noted that two features of this research are first, the objective measures studied generally require both the input and output speech data sets and, second, the tests are always performed on a actual speech data. In particular, exactly the same speech data is always used for both the objective and subjective speech quality measures.

For the purposes of this research, objective measures may be very simple, such as the traditional signal-to-noise ratio, or they can be very complex. A complex measure might use such diverse quantities as a spectral or other parametric distance between the input and output speech data sets; objectively computable distance measures specifically designed to predict subjective quality for a class of distortions; objectively computable distance measures specifically designed to predict parametric subjective quality; semantic, syntactic, or phonemic information extracted from the input speech data set; or the characteristics of a talker's vocal tract or glottis. The objective measures studied as part of this research program make no explicit use of semantic, syntactic, or phonemic information, but they do utilize all of the other classes of information listed above. If an objective measure satisfies the triangle inequality and other conditions shown in Figure 2.2-1, then it is a metric. Although metrics have many desirable properties, an objective measure need not be a metric to be of interest.

The procedure developed for the testing of objective speech quality measures is illustrated in Figures 2.2-2 and 2.2-3. Figure 2.2-2 describes the procedure for untrained objective measures, while Figure 2.2-3 describes the procedure for trained objective measures. The entire procedure is based on an input speech data set called the undistorted speech data base which in this study, consists of one set of twelve Harvard phonemically balanced sentences,

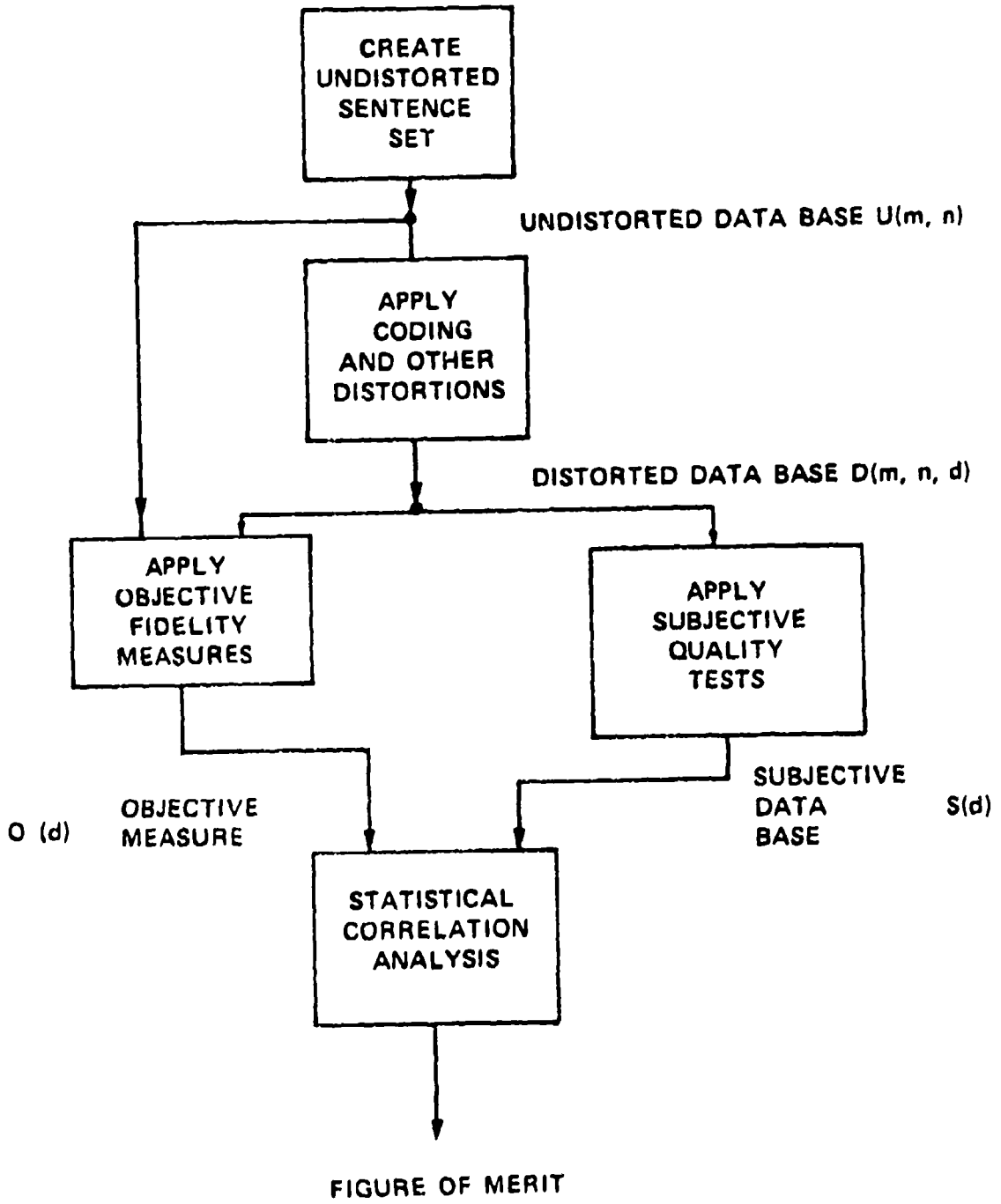


Figure 2.2-2. Block Diagram for System for Comparing the Effectiveness of Objective Quality Measures.

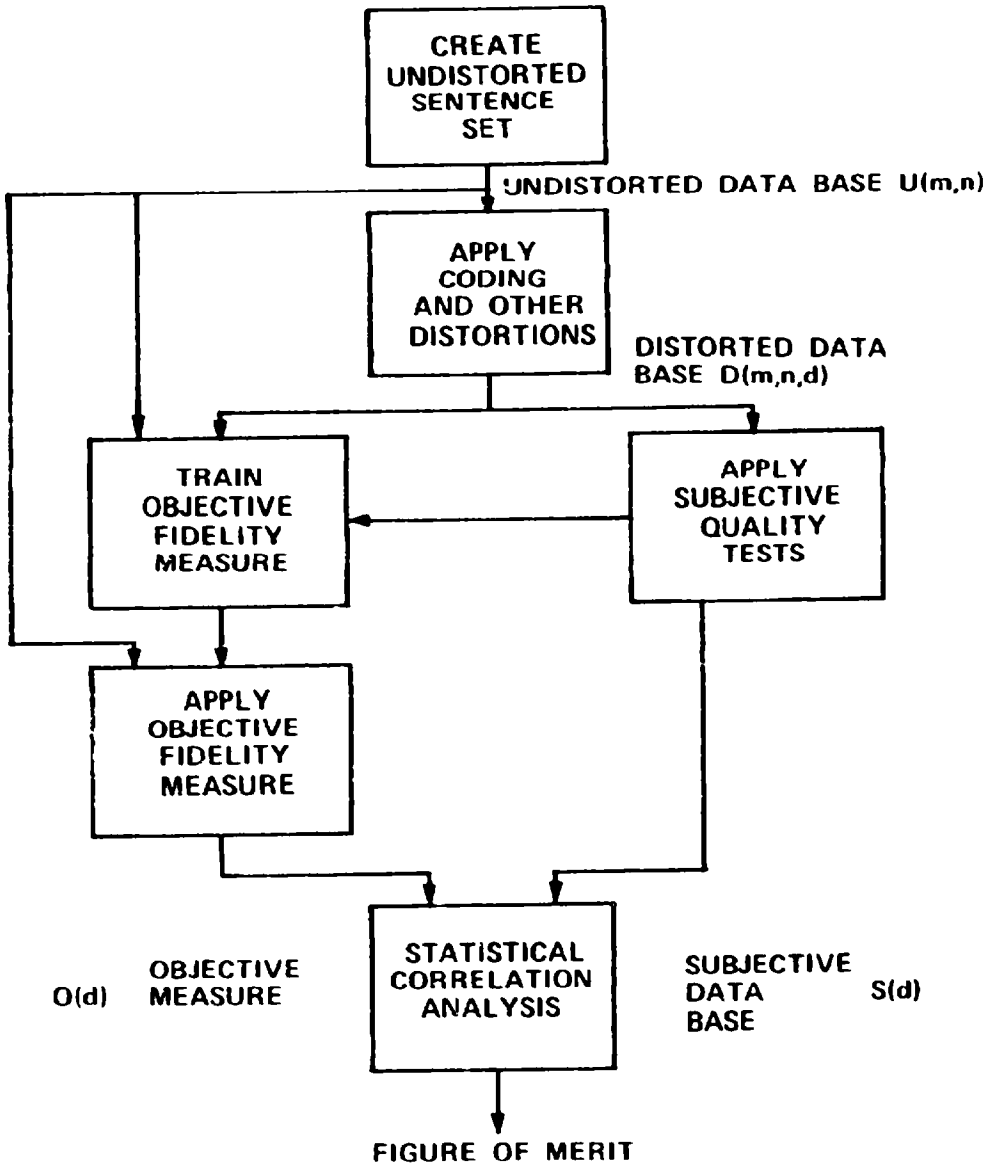


Figure 2.2-3. Block Diagram for System for Comparing the Effectiveness of Trained Objective Speech Quality Measures.

spoken by each of four talkers. The four talkers included a low-pitch male, two moderate-pitch males, and a moderate-pitch female. The 48 sentences were filtered using a tenth order elliptic lowpass filter with a 3.2 kilohertz cutoff frequency, and were sampled at an eight kilohertz rate with 12-bit A-to-D converter. This particular format was chosen so that the input speech signals would be approximately toll quality, although the speech samples were not passed through a highpass filter, as would occur for true telephone speech. The entire undistorted speech data base contained about four minutes of speech. All of the sampled speech in this study was stored on magnetic media as 16-bit integer data in digital form.

The distorted speech data base was generated by applying a large number of distortion generation (e.g., digital coding) systems to the signals in the undistorted speech data base. The distorting systems were generally implemented as FORTRAN programs designed for the network of minicomputers and array processors comprising the Georgia Tech Digital Signal Processing Laboratory [see Appendix A]. In every instance, great care was taken to synchronize the input and output speech signals at least on a frame-by-frame basis, and on a sample-by-sample basis whenever possible. This completely eliminated the problem of synchronizing the undistorted and distorted speech signals, and the synchronization problem was not addressed by this research. At the beginning of this research contract, the distorted speech data base contained speech generated by 264 distorting systems, for a total of 4x12x264=12672 sentences, or 14.42 hours of distorted speech. As part of this research, an additional 58 distorting systems were added, bringing the total to 15456 sentences, or 17.59 hours of distorted speech. The details of the pre-existing data base are described in section 2.3, while the new speech distortions are described in Chapter 3.

The third major component of the objective measure testing procedure is

the subjective data base, which is formed by applying a subjective speech quality measure to all the distortions in the distorted speech data base. In this study, the subjective test used was the Diagnostic Acceptability Measure, or DAM, developed by William D. Voiers at the Dynastat Corporation [2.1][2.6]. This is a widely used subjective quality test of the mean opinion score variety in which subjects are asked to assign a number to their perception of the quality of the speech samples under consideration, and a final system quality score is derived from these individual quality assessments. The DAM test has the great advantage that it not only gives isometric quality assessments, such as perceived acceptability or perceived pleasantness, but also gives estimates of parametric subjective qualities as well. The latter of these include such things as system fluttering, SF, or system lowpass, SL. In addition, the DAM also allows subjects to differentiate between background and foreground distortions. Details of the DAM and the subjective data base are discussed in section 2.4 and Chapter 3.

Two broad classes of objective speech quality measures which were addressed as part of this study were untrained objective measures; and trained objective measures. In the former, all the parameters which control the objective measure are fully specified as part of the definition of the objective measure itself. In the latter, some of the control parameters for the objective measures are statistically optimized using the data in the three data bases.

The untrained objective measures are tested as shown in Figure 2.2-2. First, the objective quality measure is applied to all of the distortions in the distorted speech data base, using the undistorted speech data base as reference. Second, a statistical correlation analysis is done between the results from the objective measure and corresponding results from the

subjective data base. The results from the statistical analysis are used as a figure-of-merit for comparing different objective speech measures.

Two figures-of-merit have been used throughout this research program. The first is an estimate of the correlation coefficient between the objective quality measure, O(d) (where d is the index of the distortion) and the subjective quality measure, S(d). This estimate is given by

$$\tilde{\rho} = \frac{\sum_{d} (S(d)-\bar{S}(d))(O(d)-\bar{O}(d)))}{[\sum_{d} (S(d)-\bar{S}(d))^2]^{1/2}[\sum_{d} (O(d)-\bar{O}(d))^2]^{1/2}} \qquad 2.2\text{-}1$$

This results in a minimum variance linear estimate of the subjective quantities from the objective quantities given by

$$S(d) = S(d) + \frac{\tilde{\rho}\tilde{\sigma}_s}{\tilde{\sigma}_0} (O(d)-\bar{O}(d)) \qquad 2.2\text{-}1$$

where $\tilde{\sigma}_s$ and $\tilde{\sigma}_0$ are the estimated standard deviation for the subjective and objective measures respectively. It would not be correct to attribute any absolute validity to this estimated correlation coefficient in relation to other studies. For example, since we have not randomly sampled the universe of all coding distortions, our correlation estimates are biased by the content of our distortion ensemble. Therefore, correlation estimates computed in this way are only meaningful when comparing objective measures over exactly the same distortion ensemble, and such estimates should certainly not be compared otherwise.

A more universal figure-of-merit can be computed if the objective estimate of the subjective data is viewed as a linear regression analysis. The desired figure-of-merit is the expected standard deviation of error when the subjective

results are estimated entirely from the objective results, given by

$$\tilde{\sigma}_e = [E[(S-D(S|0))^2]]^{1/2} = [\tilde{\sigma}_s^2(1 - \tilde{\rho}^2)]^{1/2} \qquad 2.2\text{-}3$$

This estimate, which incorporates the variance of the subjective data base as well as the correlation coefficient, is a more pleasing figure-of-merit since it can be viewed as an actual performance estimate.

The trained objective measures are tested as shown in Figure 2.2-3. The primary difference between the trained and the untrained measures is that the trained measures are defined using some number of unspecified parameters, whereas untrained measures are defined with all parameters specified. Trained objective measures are tested using the two-pass procedure of Figure 2.2-3. In the first pass, the regression coefficients for the objective measure under test are set so as to maximize the correlation between the objective and subjective results. Then, in the second pass, this now fully specified objective measure is tested exactly like an untrained measure. In this procedure, if the data in the training set is the same as the data in the testing set, then the figures-of-merit estimate an upper bound on the performance of the objective measure under test. If separate training and testing sets are used, then the figures-of-merit form an actual performance estimate.

2.3 The Distorted Speech Data Base

As previously discussed, the distorted speech data base is generated from the undistorted speech data base through the application of a large number of distorting systems, each of which is uniquely identified by its type of distortion and its level of distortion. In general, each type of distortion was realized with six (or sometimes twelve) levels of distortion. Whenever

possible, these levels were set to span the perceived range from barely perceivable to moderately distorted. Table 2.3-1 summarizes the distortions used in this research.

As can be seen from Table 2.3-1, some of distortions in the distorted data base already existed at the beginning of this research program, while others were generated as part of this research. The pre-existing distortions are described in detail in a previous DCA report [2.1], while the new distortions are described in Chapter 3 of this report. The purpose of this section is to briefly review the distortions which were generated as part of the previous DCA research program.

2.3.1 Coding Distortions

The purpose of the coding distortions was to include in the distorted speech ensemble a reasonable cross-section of the digital coding techniques. Those included in the original data base were chosen from among systems which were either in use or under active development in 1978. As can be seen from Table 2.3-1, these coding distortions can be roughly divided into two classes: waveform coders and vocoders. The waveform coders included six time-domain coders (ADM, CVSD, APCM, ADPCM, and APC) and one frequency domain coder (ATC). The vocoders were all based on linear predictive coding techniques, and included two voice excited (now more commonly call residual excited) vocoders (VEV) and one pitch excited vocoder (LPC).

Among the waveform coders, two different adaptive delta modulators were included in the distortion ensemble: ADM and CVSD. The ADM system, which was based on a technique proposed by Jayant [2.7] used a one-bit memory to control its quantizer adaption and one-tap linear predictor in which the predictor constant was chosen to minimize the mean square prediction error at the operating bit rates across the entire input speech set. In addition, the quantizer attack and decay rates were chosen to be equal [2.1] [2.7]. The

Coding Distortions	Number of Cases	Added During Current Study
ADPCM	6	No
APCM	6	No
CVSD	6	No
ADM	6	No
APC	6	No
LPC Vocoder	6	No
VEV	12	No
ATC-1	6	No
ATC-2	6	Yes
SBC	6	Yes
ADPCM+Noise Feedback	6	Yes
MP-LPC	6	Yes
Channel Vocoder	6	Yes
Controlled Distortions		
Additive Noise	6	No
Low Pass Filter	6	No
High Pass Filter	6	No
Band Pass Filter	6	No
Interruption	12	No
Clipping	6	No
Center Clipping	6	No
Quantization	6	No
Echo	6	No
Frequency Variant Controlled Distortion		
Additive Color Noise	36	No
Banded Pole Distortion-1	78	No
Banded Frequency Distortion	36	No
Banded Pole Distortion-2	24	Yes

Table 2.3-1 Summary of Coding and Controlled Distortions in the Distorted Data Base

system was operated at 8, 12, 16, 24, and 32 KBPS, and the uncoded speech was included in this set as the sixth distortion level.

The CVSD realization used was one which had been generated as part of a separate Defense Communications Research Program [2.8]. This CVSD had been specifically optimized for tandeming with pitch excited LPC vocoders, although no tandems were included in this study. Just as for ADM, the single predictor coefficient for each CVSD bit rate was set to match the statistics of the undistorted speech ensemble. All of the CVSD systems had a minimum step size of 10 and an expansion ratio of 166 [2.1][2.8]. The CVSD was operated at the same bit rates as the Jayant ADM above.

The only difference between the two adaptive PCM systems (APCM and ADPCM) was that ADPCM used a one-tap fixed predictor (value .92) while APCM used no predictor. Both systems used a feedback exponential quantizer adaption technique similar to the approach used in CVSD [2.1][2.8]. Both systems were operated at bit rates of 12.7, 18.6, 22.5, 25.3, 27.6, and 29.6 Kbps.

The Adaptive Predictive Coder [2.9] simulated in this study used a tenth order, time varying, linear predictor which was updated every fifteen msec. The LPC coefficients were generated using the autocorrelation method [2.10], and were quantized using inverse sine quantization [2.11]. The residual encoder was of the adaptive feed forward type, and used a three level quantizer. The APC was operated at rates of 13.3, 13.9, 14.5, 15.2, and 15.8 Kbps. The sixth distortion level used unquantized (32-bit floating point) LPC coefficients.

The adaptive transform coder (ATC), was, by modern standards, a relatively primitive transform coder. In particular, it was based on the original work by Zelinski and Noll [2.12] but used an LPC based spectral estimation procedure to assign the bits to its different channels [2.1]. This is somewhat similar to the technique later used by Tribolet and Crochiere [2.13], but without their pitch utilization technique. The LPC coefficients were also quantized, and the

transform coder was operated at rates of 20, 16, 12, 11, 9.6, and 8 Kbps.

Both of the so called voice excited vocoders (VEV) were really residual excited vocoders where only the lower frequencies of the residual signal were retained in the transmitted signal. At the synthesizer, the high frequencies in the excitation signal were regenerated using a hard-limiting operation and an additional tenth order LPC whitening filter. Like the APC and the pitch excited LPC vocoder, the VEV's used an inverse sine quantizer for the LPC coefficients. The adaptive quantizer for the decimated residual signal was of the feed-forward type, and the fundamental difference between the two VEV systems was in the rate at which the residual signal was transmitted; 5615 and 7400 bps, respectively. The first VEV operated at rates of 9.5, 8.8, 8.1, 7.5, 6.9, and 6.6 Kbps, while the second VEV operated at rates of 11.3, 10.6, 9.9 9.3, 8.7, and 8.4 Kbps.

The pitch excited LPC vocoder also used an inverse sine quantization procedure for the LPC coefficients, and a differential encoder for the pitch and gain information. The pitch detector used was of the homomorphic type, although some pitch period and voicing errors were manually corrected. This was an intentional attempt to force the primary distortion in the coder to be from the vocal tract representation and not from pitch errors. The LPC vocoder used a fifteen msec frame interval, and operated at data rates of 1.8, 2.4, 3.0, 3.7,and 4.3 Kbps. The sixth distortion level used unquantized (32-bit floating point) LPC coefficients.

2.3.2 Controlled Distortions

A large portion of the distortions generated in the original research program were not explicit coding distortions, but were controlled distortions. Each of these distortions were included for one of two reasons. Either they were considered to be examples of specific types of subjectively relevant

distortions, or they were considered to be a type of distortion which does occur in coding systems, but which does not occur in isolation.

There were fundamentally two classes of controlled distortions in the initial distorted speech data base: simple distortions; and frequency variant distortions. The frequency variant distortions were included for two main reasons. First, they could be used to measure the relative importance of different types of distortions when they are applied in different frequency bands. Second, frequency variant controlled distortions offer an environment in which frequency variant objective measures could be expected to be relatively uncorrelated between frequency bands.

Table 2.3-1 give a summary of the controlled distortions used in the original study. The simple controlled distortions included additive noise, lowpass filtering, highpass filtering, bandpass filtering, interruption, clipping, center clipping, quantization, and echo. The frequency variant distortions included additive colored noise, banded pole distortion, and banded frequency distortion.

Most of the simple controlled distortions can be described in only a few words. The additive noise, for example, was white and Gaussian, and the resulting waveforms had SNR's of 30, 24, 18, 12, 6, and 0 dB. Likewise, both the highpass and lowpass filtering distortions had cutoff frequencies of 400, 800, 1300, 1900, 2600, and 3400 Hertz. The bandpass filters had passbands of 0-400, 400-800, 800-1300, 1300-1900, 1900-2600, and 2600-3400 Hertz. It should be noted here that all of the bandpass distortions and some of the lowpass and highpass distortions were quite severe, and were unique in that regard.

The interruption distortions were implemented by multiplying the input speech signals by periodic waveforms which alternated between the values one and zero. Two different periods were used for these signals: the long period, which was 125 msec; and the short period, which was 37.5 msec. The level of

distortion for interruption was varied by changing the duty cycle of the periodic waveforms.

Both of the clipping distortions were implemented using a threshold at which the waveform was appropriately clipped. In terms of a percentage of the available dynamic range of the input speech signals, these were given by 15%, 7.6%, 3.8%, 3.05%, 1.53%, and .76% for clipping, and by 7.6%, 3.8%, 1.9%, .76%, 38%, and .19% for center clipping.

The quantization distortion was implemented as a fixed, linear PCM system which used 64, 48, 32, 24, 16, and 12 levels per sample. This corresponded to bit rates of 48, 44.7, 40, 36.7, 32, and 28.7 Kbps, respectively. Finally, the echo distortion was formed by adding a delayed version of the input speech signal back to itself. The delays used were 1.25, 6.25, 12.5, 25, 62.5, and 125 msec.

The original study included a total of three types of frequency variant distortions. The first, additive colored noise, was designed to approximate waveform coder distortions in a frequency variant way. The second, banded pole distortion, was designed to approximate distortions typical of vocal tract modeling vocoders and APC's in a frequency variant way. Finally, banded frequency distortion was designed to approximate the distortions found in ATC's and adaptive subband coders in a frequency variant way. All of the frequency variant distortions operated in six frequency bands. The band limits used were 0-400, 400-800, 800-1300, 1300-1900, 1900-2600, and 2600-3400 Hertz.

The additive colored noise was formed by first bandlimiting white Gaussian noise, and then adding the resulting signal to the original speech signals. In all, six different additive colored noise distortions were included, one for each of the frequency bands listed above. Using six distortion levels per distortion type resulted in 36 separate distorting systems.

The banded pole distortion was realized in four steps. First, an LPC analysis was performed, and a residual signal generated. Second, the LPC polynomials were factored and the pole locations were perturbed within one of the frequency bands. Third, the LPC coefficients were regenerated by multiplying together the individual perturbed poles. Finally, a distorted speech signal was generated by passing the residual signal through the regenerated LPC filter. The entire procedure is described in detail in Chapter 3 of this report. The pole perturbations were performed in both the radial and angular directions for all six frequency bands. These, plus two full-band distortions, resulted in a total of 78 separate distortions.

The banded frequency distortion was based on a short-time Fourier transform (STFT) representation for the speech signal. Fundamentally, the banded frequency distortion added noise to the STFT of the speech signal in bands. The noise was white and Gaussian, and was always added in phase with the original signal. This means that the noise was added to the magnitude of the STFT while leaving the angle undisturbed. Once again, the six frequency bands combined with six distortion levels resulted in 36 separate distortions.

2.4 The Subjective Data Base

The emphasis in this research has always been on highly intelligible coding techniques for use in toll quality applications. For this class of systems, context free intelligibility tests, such as the DRT and the MRT, are not particularly effective. This is because these high quality systems generally crowd the high end of the intelligibility scale, and hence are not well resolved by intelligibility alone. In addition, for high quality systems, it is generally acknowledged that user acceptance depends on factors other than intelligibility. The ideal type of test for this class of systems is some form of communicability test [2.16] in which a user's performance is measured on some complex or difficult task which utilizes the speech coding system

directly. Unfortunately, communicability tests are not reasonable for this research for two reasons. First, such tests are intrinsically expensive, and the cost of generating the large subjective data bases required here would be prohibitive. Second, in order to perform such tests, real-time realizations for the distorting systems are required, which would also be prohibitively expensive.

The only reasonable compromise approach left is to use a subjective preference test of the mean opinion score type. In such tests, subjects are asked to rate speech material on a subjective scale, and the distorting system's acceptability is estimated from these ratings. Subjective preference tests have the advantage that they are much less expensive to administer than communicability tests and they do not require real-time realizations for the speech distortion systems. Such tests have the disadvantage that they must deal with the subtle nature of subjective preferences and they may require the use of a large number of subjects in order to increase the test's resolving power to an acceptable level.

The subjective preference test chosen for this work was the Diagnostic Acceptability Measure (DAM) developed by the Dynastat Corporation. This particular test was chosen for several reasons. First, it is a very carefully conceived and designed measure which has been widely used and verified. Second, since it is a widely used test, its results are accepted and understood by a large number of people. Third, and most important for this research, the DAM is a very fine-grained test which measures not only such isometric subjective quantities as acceptability, but a large number of parametric quantities as well. This, in effect, generates a feature set which forms a fine-grained perceptual signature for each distortion. As will become obvious from the experimental results, without the information provided by these

parametric measures, the design of high-performance objective speech quality measures would be very difficult.

All of the Diagnostic Acceptability Measures generated as part of both the previous research program and this research program were administered by the Dynastat Corporation under subcontract to Georgia Tech.

As with most mean-opinion subjective tests, the DAM requires listeners to characterize the distorted speech in absolute, rather than relative, judgments. However, the DAM is unique in two specific ways. First, it combines the indirect parametric approach with the more conventional isometric approach, which, as previously noted, results in a much more fine-grained estimate of the speech quality. Second, the DAM allows listeners to distinguish between system and background distortion in making their judgments.

The rating form used in the DAM test is shown in Figure 2.4-1. The subjects rate the distorted speech on ten parametric system scales, seven parametric background scales, and three isometric scales. Factor analysis was previously used [2.1] to reduce the input data to the form of Figure 2.4-2. The twenty original subjective scales are reduced to fourteen output scales: six parametric system qualities (SF, SH, SD, SL, SI, and SN); four parametric background qualities (BN, BB, BF, and BR); and three isometric qualities (Intelligibility, Pleasantness, and Acceptability). From all these parameters, a total Composite Acceptability (CA) is estimated.

Previous research on the Paired Acceptability Rating Method (PARM) [2.15] has shown that much of the apparent randomness in user preference tests is actually attributable to stable differences in listener preferences. The DAM uses this fact to good advantage through the careful tracking of user performance by the use of anchors and probes. This information is then used to improve the resolving power of the DAM through the statistical correction of

DAM SYSTEM RATING FORM

Make a slash at the appropriate point on each scale to indicate the degree to which this transmission sample is characterized by the indicated quality.

THE SPEECH SIGNAL

0 10 20 30 40 50 60 70 80 90 100
FLUTTERING
Negligible TWITTERING PULSATING Extreme

0 10 20 30 40 50 60 70 80 90 100
MUFFLED
Negligible SMOTHERED LOW Extreme

0 10 20 30 40 50 60 70 80 90 100
DISTANT
Negligible SMALL COMPACT Extreme

0 10 20 30 40 50 60 70 80 90 100
RASPING
Negligible SCRAPING GRATING Extreme

0 10 20 30 40 50 60 70 80 90 100
THIN
Negligible TINNY HIGH Extreme

0 10 20 30 40 50 60 70 80 90 100
UNNATURAL
Negligible MECHANICAL LIFELESS Extreme

0 10 20 30 40 50 60 70 80 90 100
BABBLING
Negligible CHORTLING SLOBBERING Extreme

0 10 20 30 40 50 60 70 80 90 100
IRREGULAR
Negligible SPASMODIC FITFUL Extreme

0 10 20 30 40 50 60 70 80 90 100
NASAL
Negligible WHINNING DRONING Extreme

0 10 20 30 40 50 60 70 80 90 100
INTERRUPTED
Negligible INTERMITTENT CHOPPED Extreme

DAM RATING FORM (cont.)

THE BACKGROUND

0 10 20 30 40 50 60 70 80 90 100
HISSING
Negligible SIMMERING FIZZING Extreme

0 10 20 30 40 50 60 70 80 90 100
CHIRPING
Negligible CHEEPING CLICKING Extreme

0 10 20 30 40 50 60 70 80 90 100
ROARING
Negligible RUSHING GUSHING Extreme

0 10 20 30 40 50 60 70 80 90 100
CRACKLING
Negligible SCRATCHING STATICY Extreme

0 10 20 30 40 50 60 70 80 90 100
BUZZING
Negligible HUMMING WHIRRING Extreme

0 10 20 30 40 50 60 70 80 90 100
RUMBLING
Negligible THUMPING THUDDING Extreme

0 10 20 30 40 50 60 70 80 90 100
BUBBLING
Negligible GURGLING PERCOLATING Extreme

THE TOTAL EFFECT

0 10 20 30 40 50 60 70 80 90 100
INTELLIBIBLE
Negligible UNDERSTANDABLE MEANINGFUL Extreme

0 10 20 30 40 50 60 70 80 90 100
PLEASANT
Negligible RICH MELLOW Extreme

0 10 20 30 40 50 60 70 80 90 100
Negligible ACCEPTABLE Extreme

Figure 2.4-1. DAM Rating Form.

Figure 2.4-2. STRUCTURE OF THE DAM

Signal Quality Measures

SF	1,7	Fluttering Bubbling	Amplitude-Modulated Speech
SH	3,5	Distant Thin	Highpassed Speech
SD	4,14	Rasping Crackling	Peak Clipped Speech, Quantized Speech
SL	2	Muffled Smothered	Lowpassed Speech
SI	8,10	Irregular Interrupted	Interrupted Speech
SN	0	Nasal Whining	Bandpassed Speech Vocoded Speech

Background Quality Measures

BN	11,13	Hissing Rushing	Guassian Noise
BB	15	Buzzing Humming	60-120 Hz Hum
BF	12,17	Chirping Bubbling	Errors in narrow band systems
BR	16	Rumbling Thumping	Low frequency noise

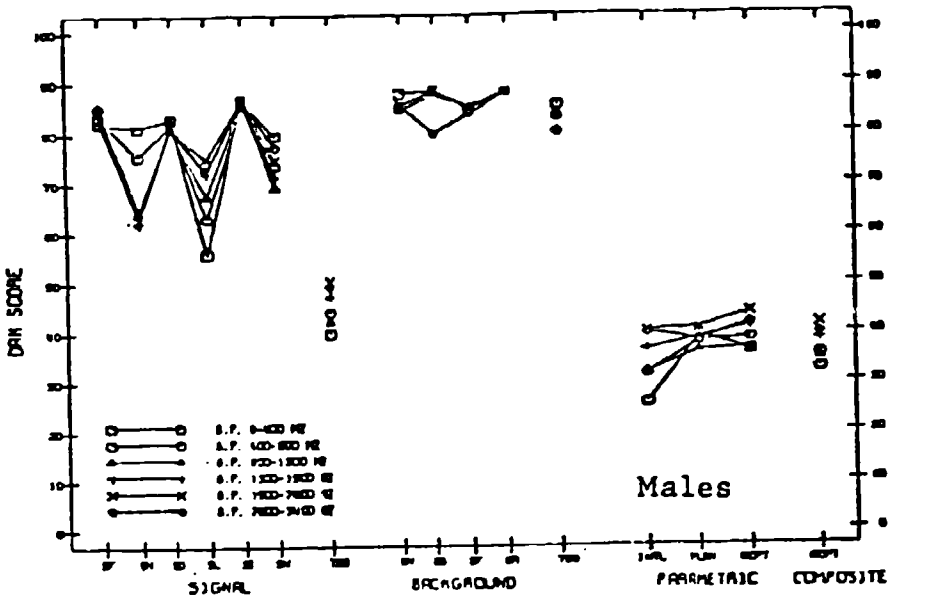
Total Quality Measures

Quality	Rating Scales Used	Representative Descriptors	Exemplars
Intelligibility	13	Intelligible	Undegraded Speech
Pleasantness	19	Pleasant	Undegraded Speech
Acceptability	20	Acceptable	Undergraded Speech

user responses. The total DAM output for a single type of distortion is illustrated in Figure 2.4-3.

At the beginning of this research program, the subjective speech data base contained the complete DAM results for the 1056 talker-distortion combinations in the initial distorted speech data base [2.1]. As the result of this research, an additional 232 combinations were added. A fairly detailed discussion of the initial subjective data base was included in the previous research report, and the interested reader is referred there for detailed information [2.1].

On the whole, it is a fair statement that the original subjective data base met its design goals. That is to say that it excited the appropriate range of perceived distortions, it excited all of the various parametric scales, and it represented a reasonable ensemble of coding distortions for the time at which it was designed (1978). There were a few specific exceptions to this statement, however. For example, a few of the controlled distortions could be characterized as severe rather than moderate. These included most of the bandpass distortions and some of the highpass and lowpass distortions. In addition, although the banded pole distortion generated subjective scores in the correct range, the spread of the distortion levels was not really wide enough. This result will be discussed more fully in Chapter 3. Many of the detailed features of the subjective data base will also be discussed in Chapter 4, Chapter 5 and Chapter 6.



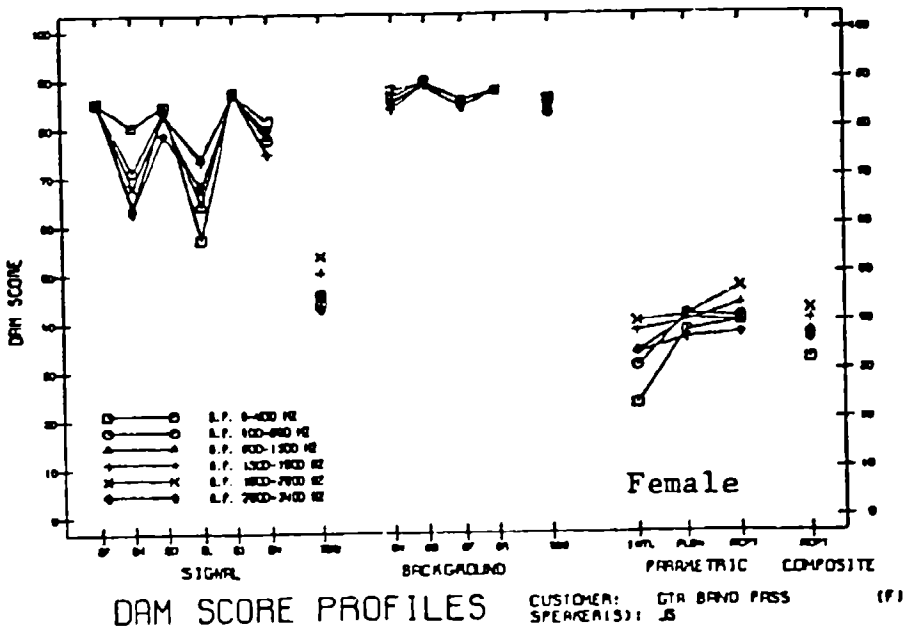


Figure 2.4-3. Effects of band-pass filtering on DAM scores for male and female speakers

REFERENCES

[2.1] T.P. Barnwell and P. Breitkopf, 'Segmental Preclassification for Improved Objective Speech Quality Measures,' Proc. of ICASSP '81, March 1981.

[2.2] T.P. Barnwell, 'Correlation Analysis of Subjective and Objective Measures for Speech Quality,' 1980 International Conference on Acoustics, Speech, and Signal Processing, Denver, Colorado, April 1980.

[2.3] T.P. Barnwell, 'A Comparison of Parametrically Different Objective Speech Quality Measures Using Analysis with Subjective Quality Results,' 1980 International Conference on Acoustics, Speech, and Signal Processing, Denver, Colorado, April 1980.

[2.4] T.P. Barnwell and W.D. Voiers, 'An Analysis of Objective Measures for User Acceptance of Voice Communications Systems,' Final Report to the Defense Communications Agency, DCA100-78-C-0003, September 1979.

[2.5] T.P. Barnwell, III, and S.R. Quackenbush, 'An Analysis of Objectively Computable Measures for Speech Quality Testing,' Proc. of ICASSP '82, May 1982.

[2.6] W.D. Voiers, 'Diagnostic Acceptability Measure for Speech,' 1977 International Conference on Acoustics, Speech, and Signal Processing, Hartford, CN, May, 1977.

[2.7] N.S. Jayant, 'Adaptive Delta Modulation with a One Bit Memory,' Bell Systems Tech. J., Vol. 49, March 1970.

[2.8] T.P. Barnwell, R.W. Schafer and A.M. Bush, 'Tandem Interconnection of LPC and CVSD Digital Speech Coders,' Final Report, DCA, DCEC, DCA 160-76-c-0073, November 1977.

[2.9] B.S. Atal and S.L. Hanauer, 'Speech Analysis and Synthesis by Linear Prediction of the Speech Wave,' JASA, 50, 1971.

[2.10] J. Makhoul, πLinear Prediction: A Tutorial Review,π Proc. IEEE , vol. 63, 1975.

[2.11] A.H. Gray, Jr. and J.D. Markel, 'Quantization and Bit Allocation in Speech Processing,' IEEE Trans. ASSP, Vol 24, No. 6, December 1976.

[2.12] R. Zelinski and P. Noll, 'Adaptive Transform Coding of Speech Signals,' IEEE Tran. ASSP, Vol. ASSP-25, No.4, August 1977.

[2.13] J.M. Tribolet and R.E. Crochiere, 'An Analysis/Synthesis Frame-Work for Transform Coding of Speech,' Proc. ICASSP-79, Washington, DC, April 1979.

[2.14] T.P. Barnwell III, A.M. Bush, R.M. Mersereau, and R.W. Schafer, 'Speech Quality Measurement,' Final Report, DCA Contract No. RADC-TR-78-122, June 1977.

[2.15] W.D. Voiers et al., 'Methods of Predicting User Acceptance of Voice Communications Systems,' Final Report, DCA 100-74-C-0056, DCA, DCEC, Reston, VA, July 1976.

CHAPTER 3

NEW SPEECH DISTORTIONS

The purpose of this chapter is to describe the new coding distortions which were added to the distorted speech data base as part of this research program. As discussed in the previous chapter, the distorted speech data base is a major component in the procedure for designing and testing the new objective speech quality measures. In general, this data base is formed by applying coding and controlled distortions to all of the sentences in the undistorted speech data base. The undistorted speech data base contains a total of four sets of twelve sentences, where the sentences were all drawn from a set of phonemically balanced sentences. Since the emphasis in this study was on communications systems which, at a minimum, come close to achieving toll quality, the undistorted sentence sets were digitized at the toll quality standard. In other words, the sentences were all band-limited to 3.2 kilohertz, sampled at eight kilohertz, and quantized to twelve bits (linear) resolution. In addition, the timing of the sentences within the sentence sets was constrained so that the distorted speech could be used directly as input for the Diagnostic Acceptability Measure (see Chapter 2 for more details). Hence, both the subjective quality estimates and the objective quality estimates in the study were always performed on exactly the same speech data.

All of the distorting systems generated as part of this study were implemented as programs (usually in FORTRAN) on the network of general purpose computers and array processors which forms the Georgia Tech Digital Signal Processing Laboratory [Appendix A]. As was discussed in Chapter 2, the distorting systems were implemented so as to maintain either sample-level or frame-level synchronization between the undistorted input speech and the distorted output speech. Hence, the problem of synchronizing the distorted and

undistorted speech was entirely avoided, and that problem was not addressed as part of this research. Both the distorted and undistorted speech sentence sets were always stored as sixteen bit integer data in disk or tape files.

The original distorted speech data base which was available at the beginning of this research effort [3.1] was described Section 2.3. In all, this data base included 264 distorting systems applied to twelve sentences for each of four talkers, for a total of 4 X 12 X 264 = 12672 sentences. The sentences are always presented at exactly 4.096 second intervals, resulting in a total distorted speech data base of 14.418 hours of distorted speech.

Fundamentally, the distorted speech data base forms the ensemble of distortions over which the statistical estimations used in the design and testing of the objective speech quality measures are performed (see Chapter 2 for more details). In an ideal statistical sense, these distortions should be a randomly selected sample from the set of all coding distortions. This, of course, is a meaningless statement for all practical applications, since clearly there exists no reasonable procedures for approaching this ideal. What was done instead was to design a distortion ensemble which is representative of the particular communications environments of interest.

The distortion ensemble in the original study was generated to conform to several specific design criteria. First, since the interest of the Defense Communications Agency is primarily in medium-to-high quality speech communications systems, all of the distortions were designed to span the perceptual range from barely perceivable to moderately distorted. In particular, the distortions included primarily systems of high intelligibility whose quality differences are most appropriately measured by mean-opinion speech quality tests such as the DAM. Second, since the final goal has always been to find objective speech quality measures to be used in conjunction with

speech coding systems, a number of coding systems were included in the distortion ensemble. In the original distorted speech data base, these were primarily representatives of the speech coding systems of interest in the 1978 time frame (see Table 2.3-1). Finally, since it is obvious that in order to design good objective speech quality measures, the fundamental mechanisms of speech perception must be addressed, a number of wide-band and frequency-variant controlled distortions were also included. For more detailed descriptions of all these distortions, the reader is referred to the previous DCA report (DA100-78-C-0003) [3.1] and to [3.2-3.13].

It is important to understand that, from a statistical viewpoint, all of the estimates performed using the distortion ensemble are biased by the procedures used in choosing the representative distortions. Stated another way, all of the results of this research must be viewed as estimates of the performance of the objective speech quality measures when operating over the distortion universe which is represented by the distortion ensemble. Hence, the validity of the results are fundamentally limited by the choice of distortions. By any measure, the data bases involved in this study are large (probably the largest available anywhere), and their associated statistical resolving power is correspondingly high. Nevertheless, they are still not nearly large enough to support a claim of universal validity.

The purpose of this chapter is to describe in detail the augmentations to the distorted speech data base which were performed as part of this research project. These additions were motivated by two problems with the existing data base. First, the results of the DAM tests which were performed as part of the original study indicated some deficiencies with certain of the frequency variant controlled distortions, specifically with the Banded Pole Distortions. Second, since 1978 a number of new and important speech coding techniques have been introduced, and these new coding distortions needed to be included in the

distorted speech data base in order to maintain the validity of the ensemble.

3.1 Banded Pole Distortion

Over the past decade, linear predictive analysis as become one of the dominant techniques in speech coding. This technique has been used in many different coding systems operating at many different bit rates. These coding systems include the pitch-excited LPC vocoder , the vector-quantized pitch-excited LPC vocoder, the residual-excited LPC vocoder, the Adaptive Predictive Coder, the Multi-pulse excited LPC vocoder, the Adaptive Transform Coder, and many more. All of these systems have the common feature that, as part of the speech coding procedure, they quantize and transmit frames of LPC coefficients in some form. In all systems where this is done, this quantization causes distortion and is perceived as distortion by listeners.

Because the quantization of LPC coefficients is such a common feature in modern speech coding systems, it is clear that understanding how to correctly predict subjective responses to this class of distortion must be one of the primary goals of this research. The problem is that the relation between LPC quantization distortion and human perception is not a simple one. LPC quantization techniques generally quantize some transformed parameter set derivable from the LPC feedback coefficients, such as the inverse-sine transformed PARCOR coefficients, the log area ratios, or the line spectral pairs. Such distortions are not frequency localized and are generally spread over the entire frequency range of the signal. Human hearing, on the other hand, is a frequency variant phenomena and responds primarily to frequency-localized and time-localized events. When viewed in the frequency domain, LPC quantization has the effect on moving the roots of the LPC polynomial, and hence the poles of the LPC vocal tract transfer function, in both bandwidth and frequency. Small variations in frequency, though easily perceivable, have

little impact on the level of perceived distortion. Bandwidth variations, however, can have dramatic perceptual effects. Bandwidths which are too narrow cause clearly perceivable 'chirps,' while bandwidths which are too large cause the speech to sound 'muffled.

In actual coding systems, the LPC coefficient quantization distortions always encompass the entire frequency range and always occur in conjunction with other classes of distortion as well. If the perceptual effects of this distortion are to be well understood, then controlled distortions need to be generated which present the LPC quantization distortion in isolation and in a frequency variant way. In the previous DCA research, the distorting system shown in Figure 3.1-1 was used to generate the pole distortion. In this system, the speech is first pre-emphasized using a second order filter, and then a framed LPC analysis is performed. The results of the LPC analysis is then used to inverse filter the original speech, giving an approximation of the glottal wave excitation [3.3].

Following the inverse filtering operation, the poles of the vocal tract function are then found by factoring the LPC polynomial. Then the banded pole distortion is applied by first identifying all the poles within a fixed frequency range, and then moving the poles slightly in either frequency or bandwidth, or both. This 'jittering' of the poles is controlled by two uniform random number generators. The 'frequency range,' FR, factor gives the range of frequency, in Hertz, in which the poles are allowed to move. The 'bandwidth factor,' BF, is a multiplicative factor controlling the bandwidth motion by

$$\text{distorted radius} = (\text{undistorted radius})[1+(BF)r] \qquad 3.1\text{-}1$$

where r is a uniform random number which ranges between plus one and minus one. Once the pole locations are distorted, they are recombined to form a new set of LPC coefficients, a'(k). These coefficients are then used to implement a new

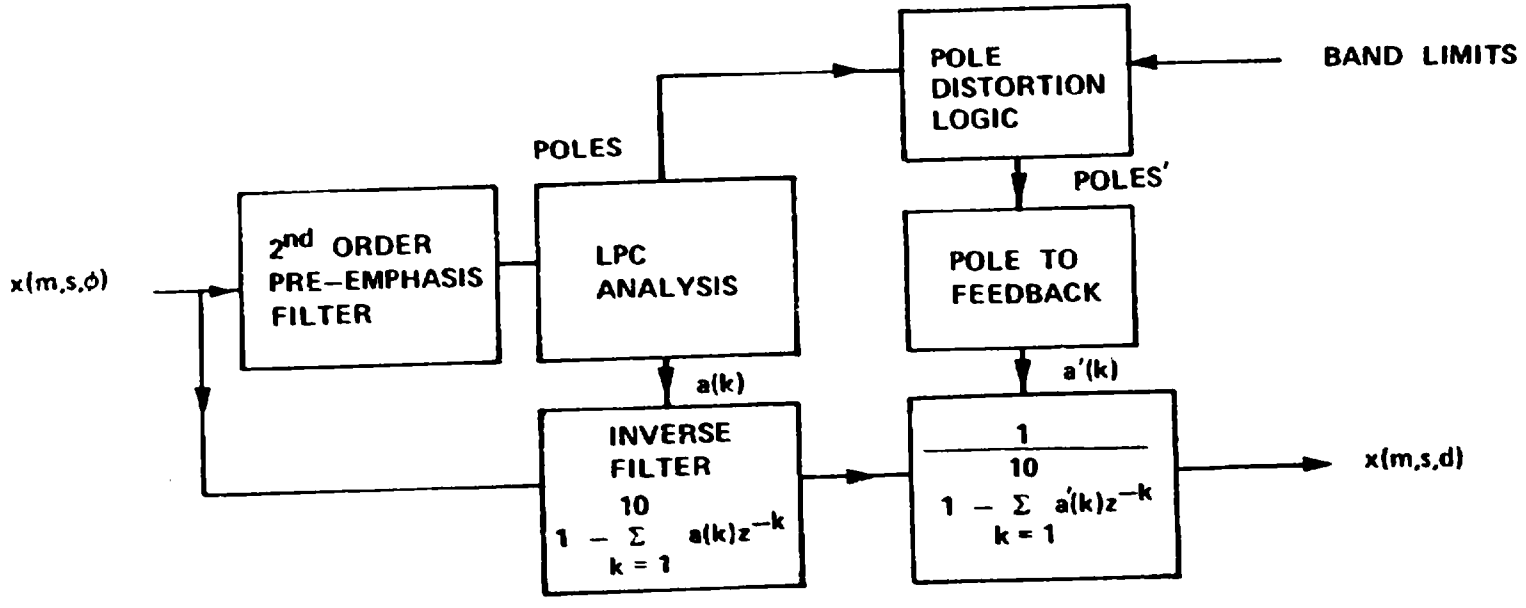


Figure 3.3-1. System for Producing the Frequency Variant Pole Distortions.

vocal tract filter to create the distorted speech. The pole distortions included in the original distortion ensemble are summarized in Table 3.1-1 and the results of the DAM analysis of these distortions are shown in Figures 3.1-2M, 3.1-2F, 3.1-3M, and 3.1-3F.

A study of the DAM results shown in Figures 3.1-2M - 3.1-3F reveals some basic problems with the distortions used in the original study. The problem is that certain of the distortion classes did not exhibit an adequate variation in perceived distortion. This is particularly true for the case of frequency distortion in the ranges 200-400 Hz, 1900-2600 Hz, and 2600-3400 Hz, but is also true for radial distortion in the range of 2600-3400 Hz. An examination of the control parameters for the banded pole distortion shown in Table 3.1-1 indicates that this is a fundamental problem, since the frequency variations used were already very large when compared to the dimensions of the frequency bands. In short, the bands used were too narrow for clearly perceivable distortions are to be generated.

Based on these observations, a new set of banded pole distortions, based on only four bands, was generated. As before, the bands were chosen to have approximately equal frequency content on a MEL scale. The control parameters for this study are shown in Table 3.1-2. Notice that in this study, the banded pole distortions were chosen so as to exhibit both pole-frequency and pole-bandwidth variations. The results of the DAM tests applied to these distortions will be discussed in the following section.

3.2 Effects of Banded Pole Distortions on Subjective Responses

Figures 3.2-1, 3.2-2, 3.2-3, and 3.2-4 show the effect of frequency variant pole distortion for 0-420 Hz., 420-900 Hz., 900-1600 Hz., and 1600-3200 Hz. respectively. From these figures, it is clear that, for all frequency ranges, the scales which are most dramatically effected are SF (system fluttering) and BF (background fluttering). Hence, the effect of quantizing

Banded Pole Distortion
Frequency Distortion

Distortion Band (Hertz)	Frequency Range (Hertz) 1	2	3	4	5	6
200-400	20	40	60	80	100	120
400-800	20	40	60	80	100	120
800-1300	50	90	130	170	210	250
1300-1900	50	90	130	170	210	250
1900-2600	100	150	200	250	300	250
2600-3400	150	200	250	300	350	400

Bandwidth Distortion

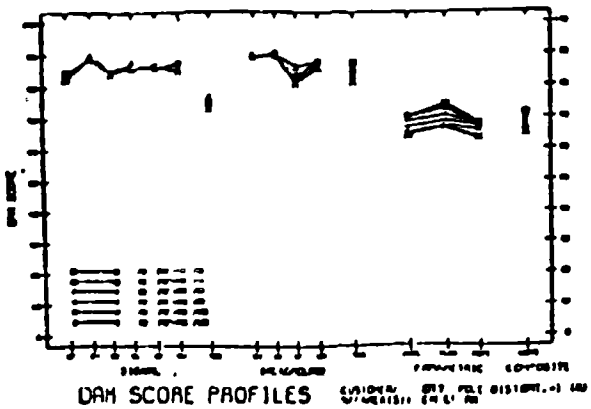
Distortion Band (Hertz)	Variation Factor 1	2	3	4	5	6
0-400	.025	.05	.075	.1	.2	.3
400-800	.025	.05	.075	.1	.2	.3
800-1300	.025	.05	.075	.1	.2	.3
1300-1900	.025	.05	.075	.1	.2	.3
1900-2600	.025	.05	.075	.1	.2	.3
2600-3400	.025	.05	.075	.1	.2	.3

Table 3.1-1 Summary of Control Parameters for the Banded Pole Distortions Implemented as Part of the Original Research

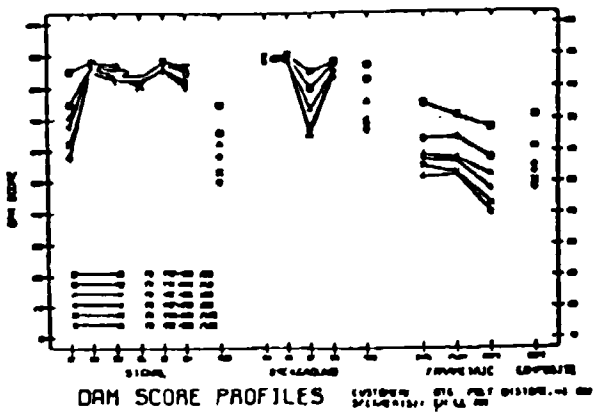
Banded Pole Distortion

Distortion Band (Hertz)	Frequency Range (Hertz) 1	2	3	4	5	6	Variation Factor 1	2	3	4	5	6
50-420	10	20	30	40	50	55	.01	.02	.04	.08	.16	.32
420-900	20	40	60	80	100	120	.01	.02	.04	.08	.16	.32
900-1600	25	50	75	100	125	150	.01	.02	.04	.08	.16	.32
1600-3200	80	160	240	320	400	500	.01	.02	.04	.08	.16	.32

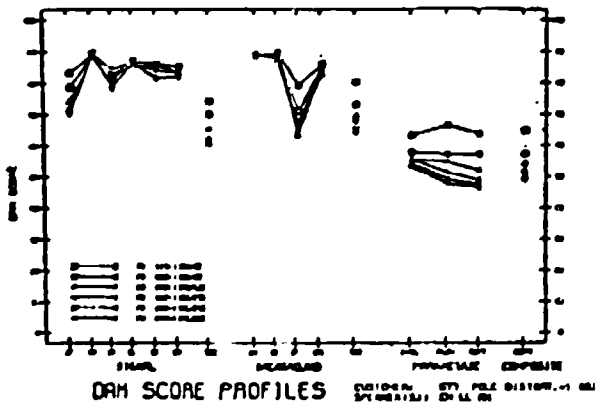
Table 3.1-2 Summary of Control Parameters for the Banded Pole Distortions Implemented as Part of the Current Research



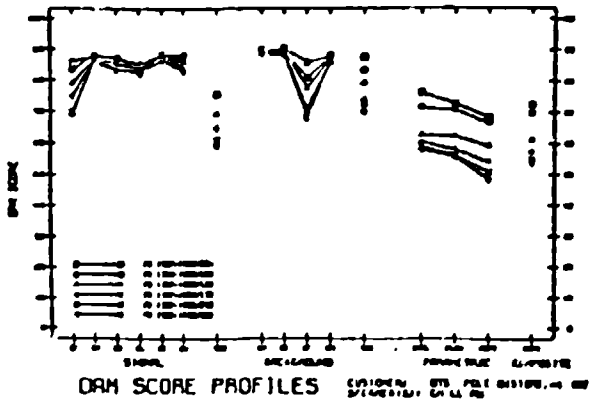
Band 1: 200-400 Hz



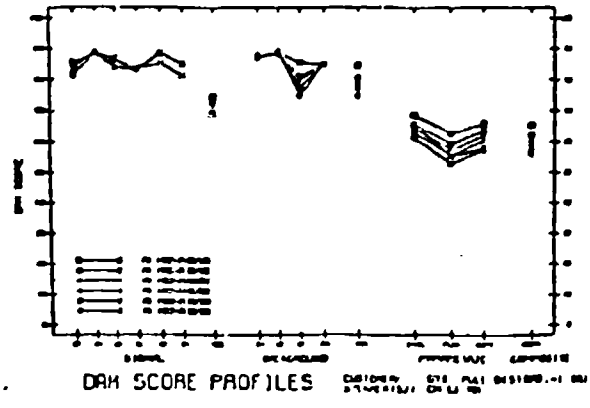
Band 2: 400-800 Hz



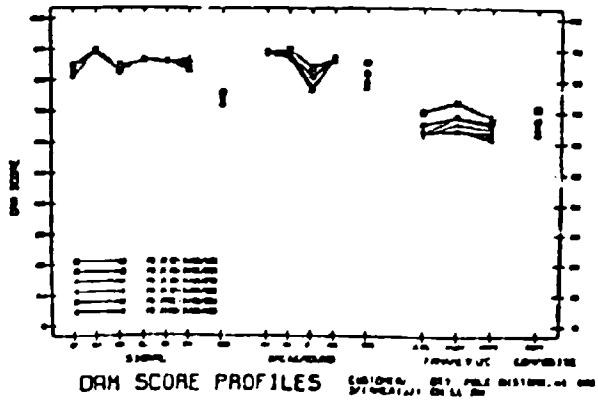
Band 3: 800-1300 Hz



Band 4: 1300-1900 Hz

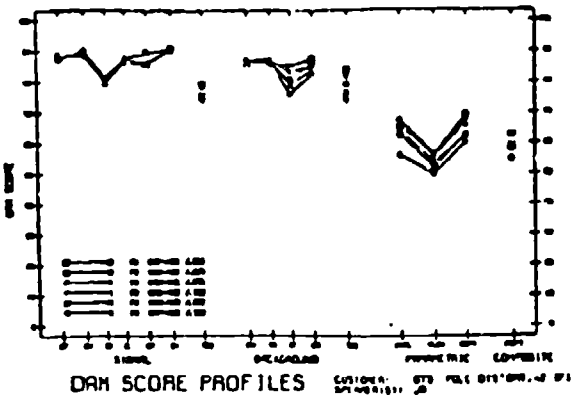


Band 5: 1900-2600 Hz

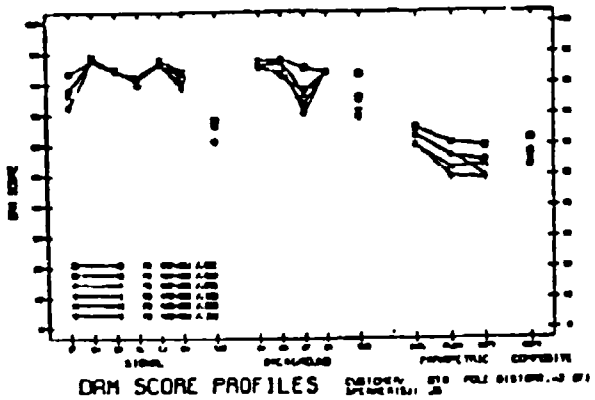


Band 5: 2600-3400 Hz

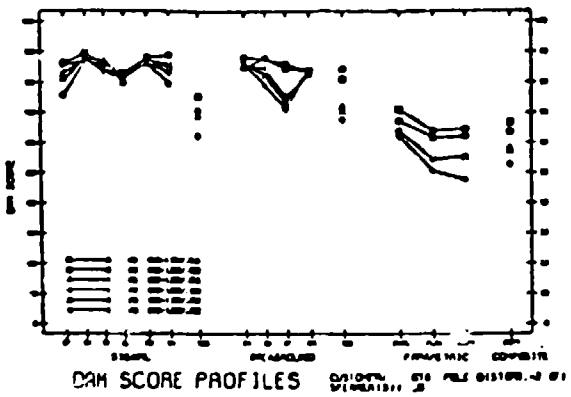
Figure 3.1-2M Effects of pole-frequency distortion on DAM scores for male speakers.



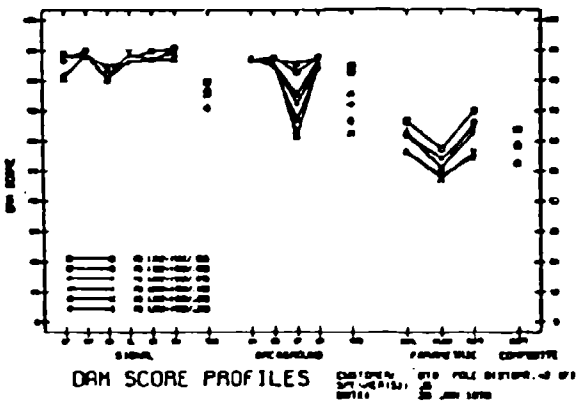
Band 1: 0-400 Hz



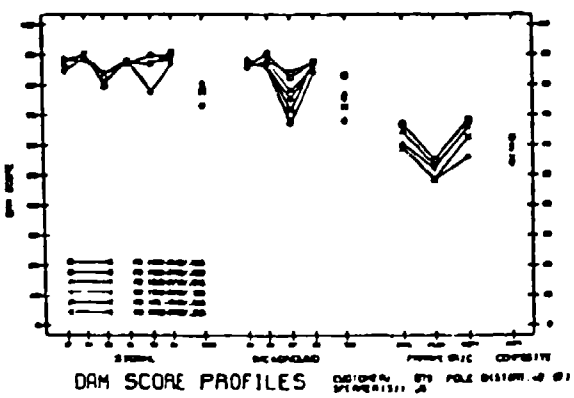
Band 2: 400-800 Hz



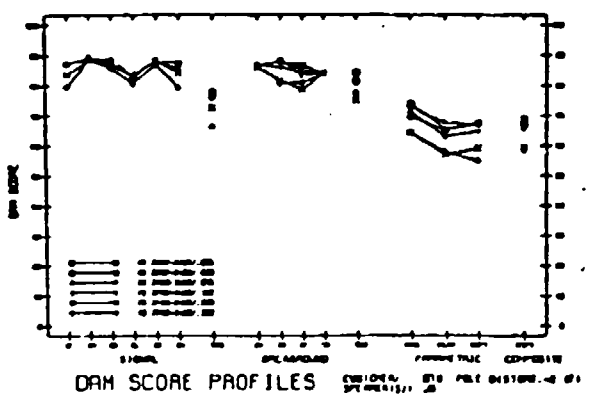
Band 3: 800-1300 Hz



Band 4: 1300-1900 Hz

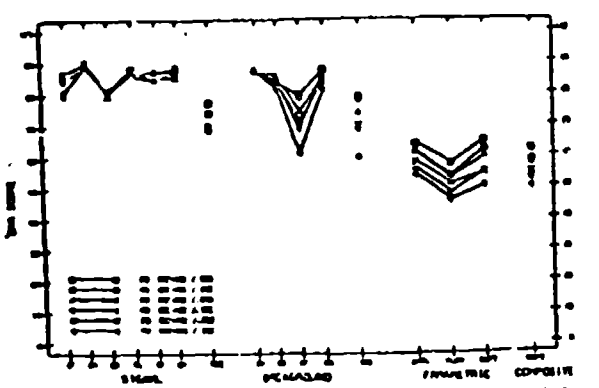


Band 5: 1900-2600 Hz

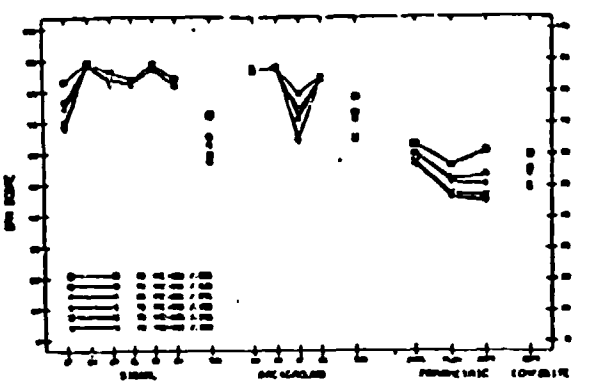


Band: 2600-3400 Hz

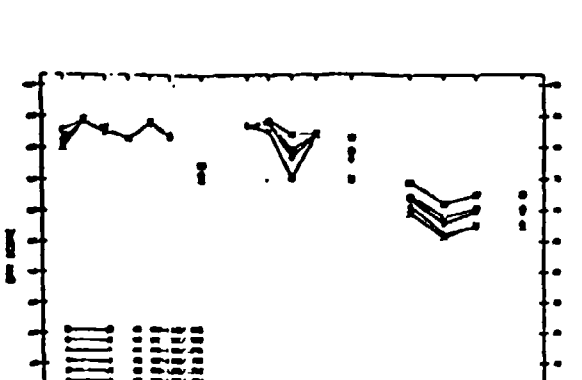
Figure 3.1-2F Effects of radial pole distortion on DAM scores for female speaker.



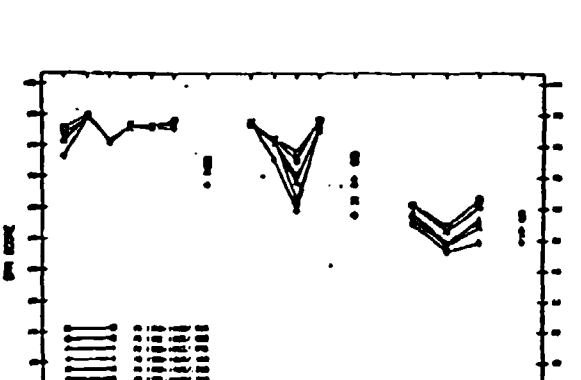
Band 1: 0-400 Hz



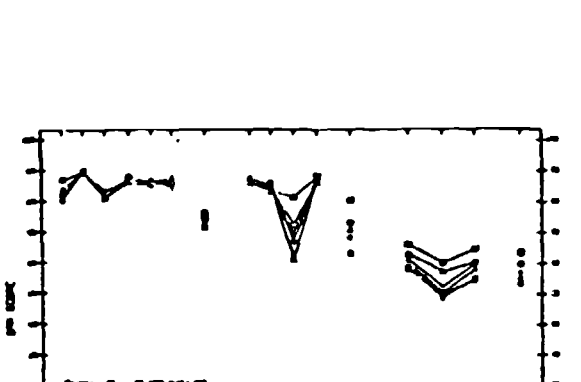
Band 2: 400-800 Hz



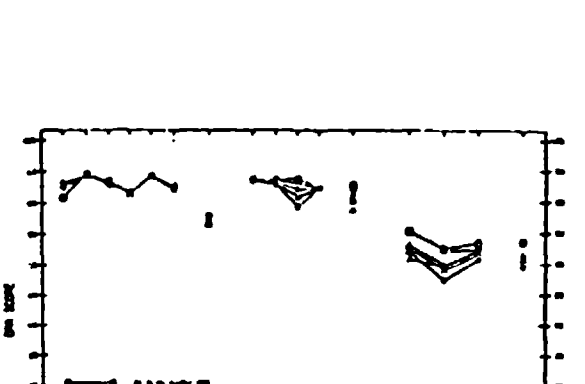
Band 3: 800-1300 Hz



Band 4: 1300-1900 Hz

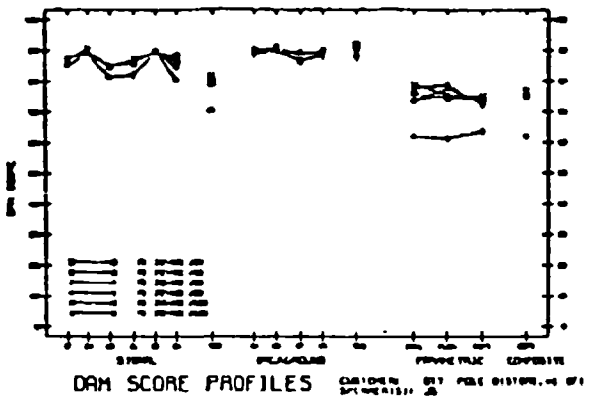


Band 5: 1900-2600 Hz

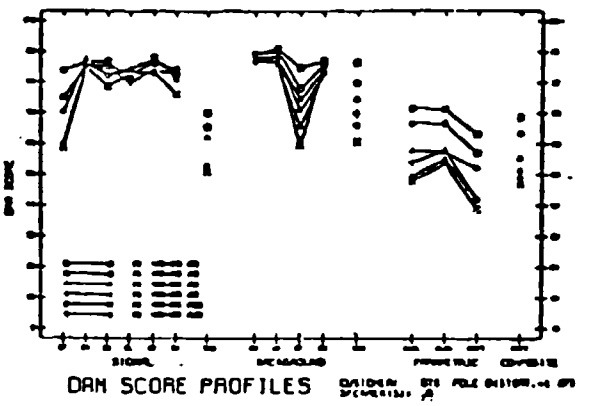


Band 6: 2600-3400 Hz

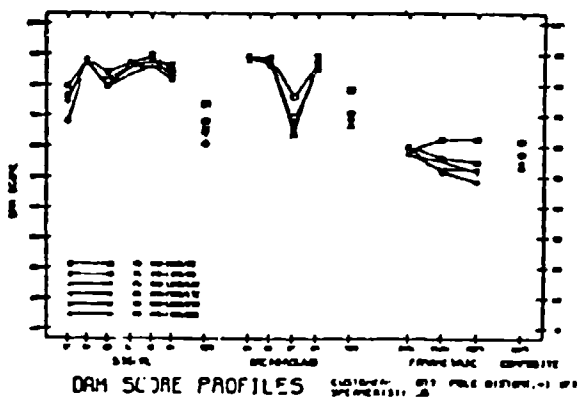
Figure 3.1-3M Effects of radial pole distortion on DAM scores for male speakers.



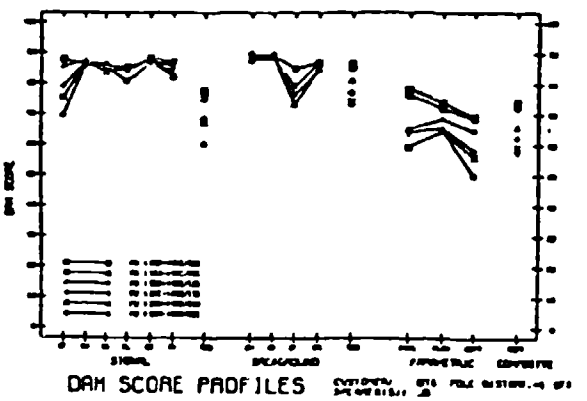
Band 1: 200-400 Hz



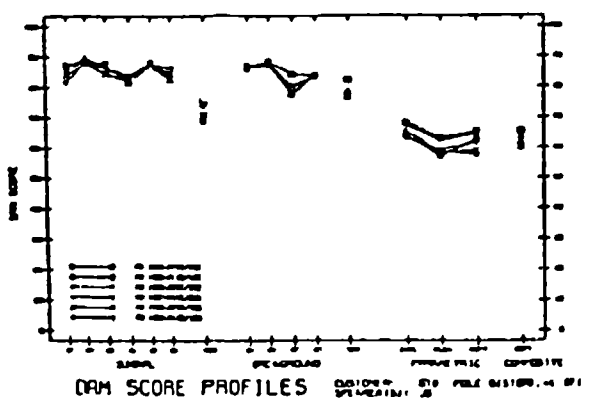
Band 2: 400-800 Hz



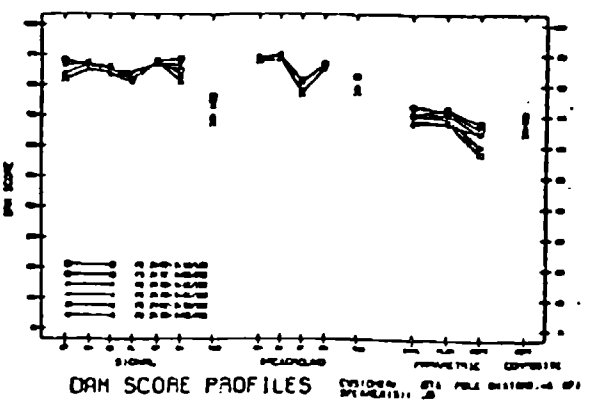
Band 3: 800-1300 Hz



Band 4: 1300-1900 Hz



Band 5: 1900-2600 Hz



Band 6: 2600-3400 Hz

Figure 3.1-3F Effects of pole-frequency distortion on DAM scores for female speaker.

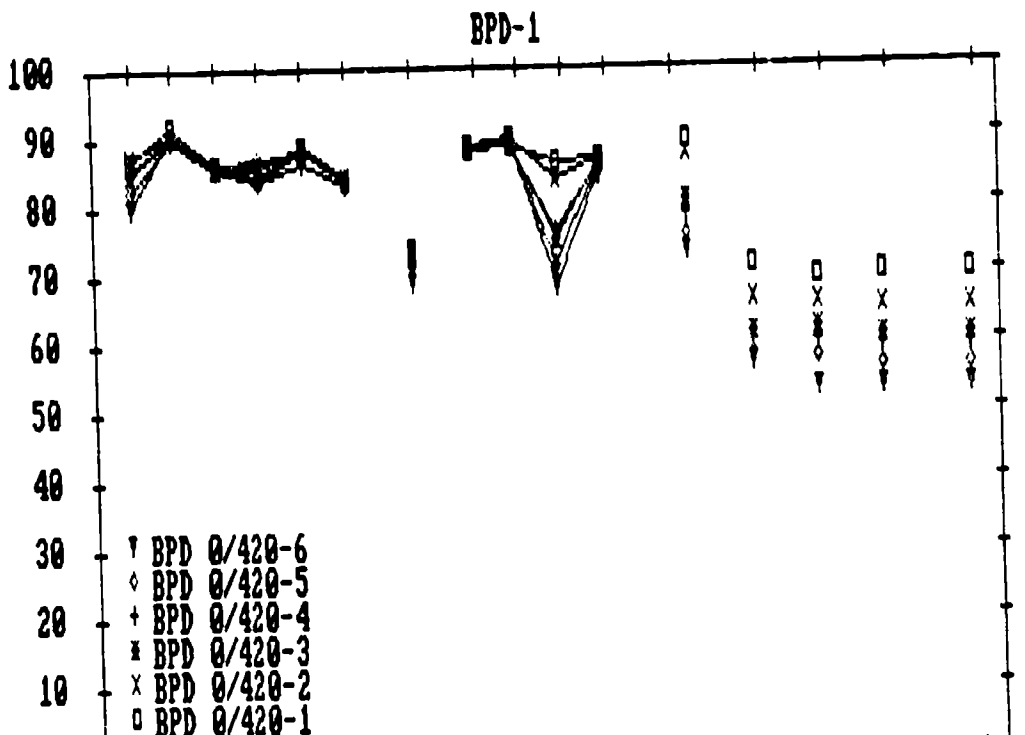


Figure 3.2-1 Diagnostic Acceptability, Measure Results for New Banded Pole Distortions in the 0-420 Hertz Frequency Band.

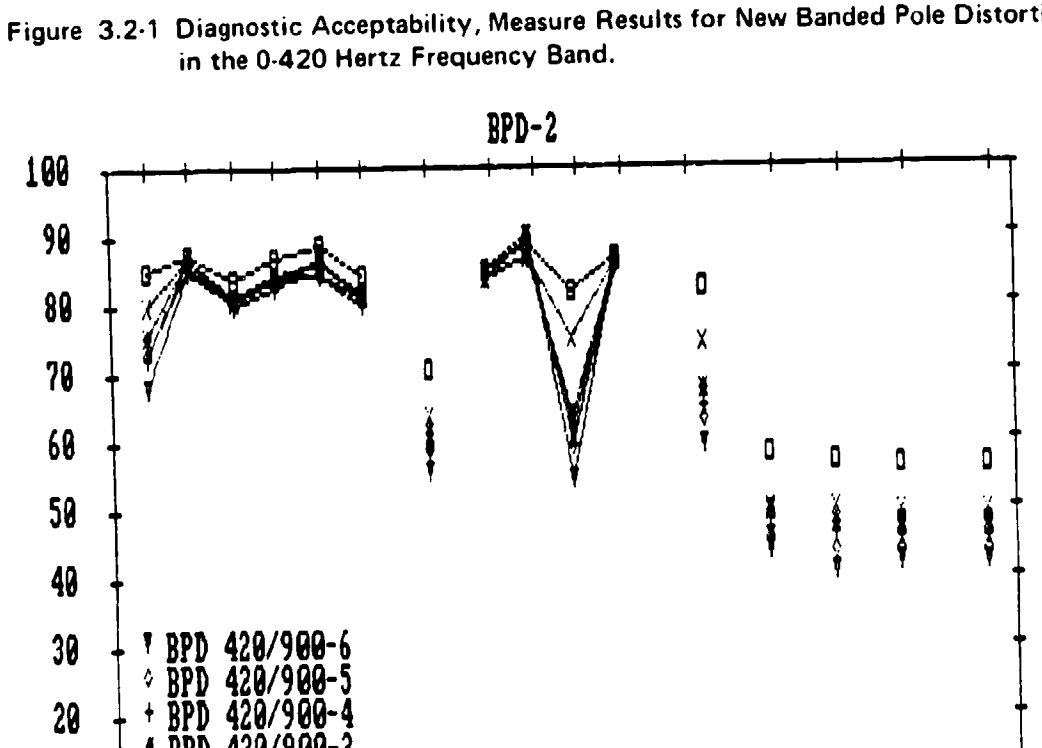


Figure 3.2-2 Diagnostic Acceptability, Measure Results for New Banded Pole Distortions in the 420-900 Hertz Frequency Band.

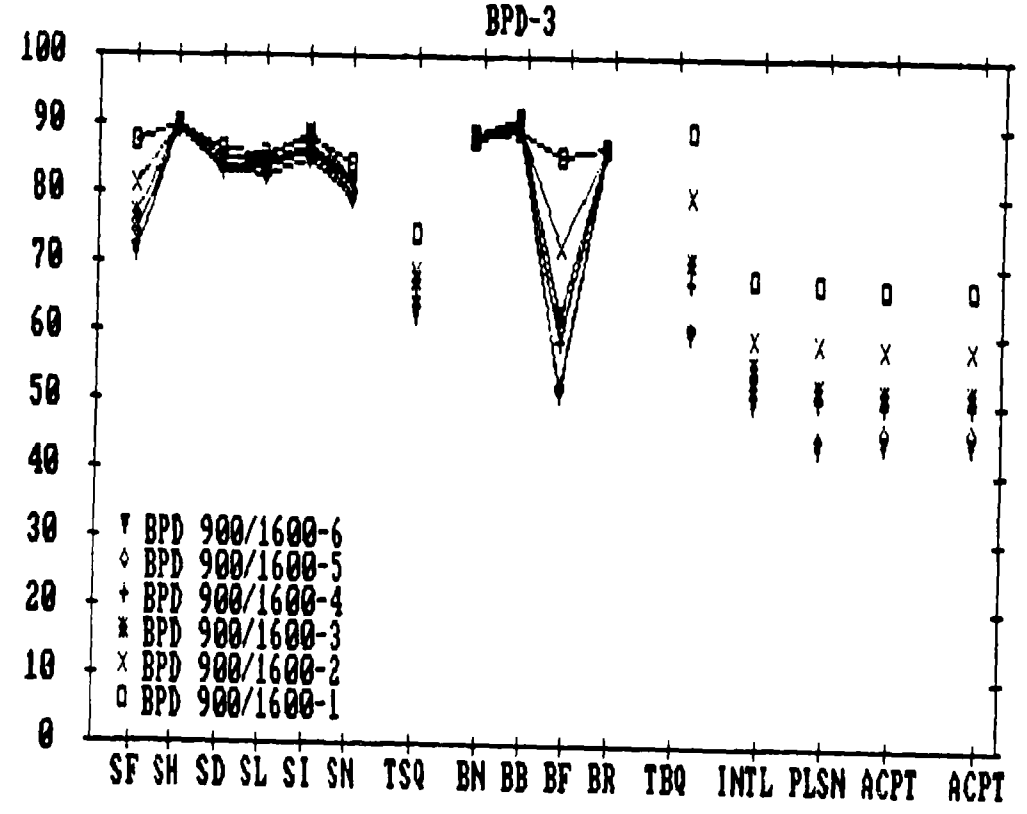


Figure 3.2-3 Diagnostic Acceptability Measure Results for New Banded Pole Distortions in the 900-160 Hertz Frequency Band.

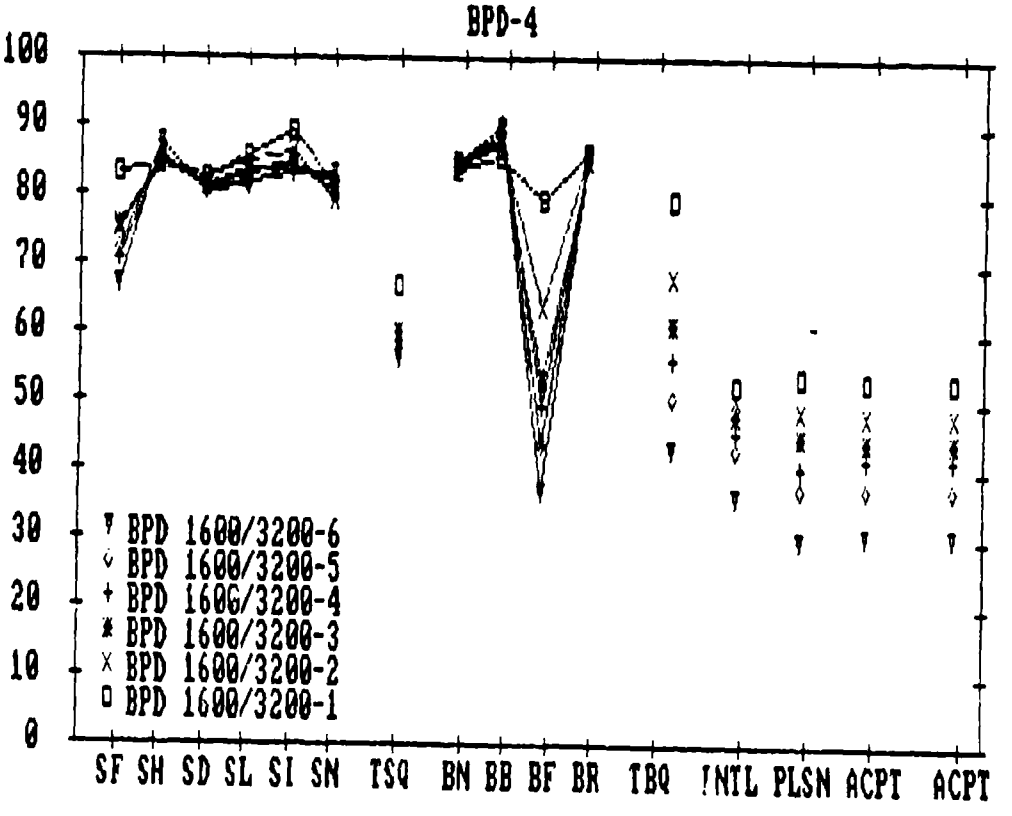


Figure 3.2-4 Diagnostic Acceptability Measure Results for New Banded Pole Distortions in the 1600-3200 Hertz Frequency Band.

the pole locations for LPC analysis can best be characterized as 'fluttering' and 'chirping'. It is also clear that all frequency bands result in an acceptably wide range of perceived distortions. Hence, the new pole distortions met their fundamental design criteria.

3.3 Coding Distortions

As previously noted, the basic reason for the introduction of new coding distortions into the distorted speech data base was to add to the distortion ensemble examples of classes of coding distortions which have become common since the original definition of the data bases in 1978. In all, there were five new classes of coding distortions introduced, resulting in a total of 34 new distortions and extending to 94 the total number of coding distortions in the distorted speech data base. As always, the new coding distortions were simulated using general purpose computers, and were designed to have zero phase reconstruction whenever possible. If this was not possible, they were designed to have at least frame-by-frame synchronization with the undistorted speech.

3.3.1 Multi-Pulse Linear Predictive Coder

Since its introduction in 1981 [3.14], the Multi-pulse Linear Predictive Coder (MPLPC) has been one of the most extensively reported and studied [3.15-3.17]] techniques for medium-to-low bit rate speech coding. For nearly a decade before 1981, researchers had been searching for ways to improve the quality of speech at the bit rates between the medium-bit-rate waveform coders (down to about 16 Kbps) and the low-bit-rate pitch-excited vocoders (down to about 2.4 Kbps), but little progress had been made. MPLPC is the first technique to show real promise in this area.

MPLPC is really a form of residual excited vocoder where the excitation information is generated and encoded in a special way. MPLPC derives its advantage from extensive utilization of the speech model and the LPC-estimated vocal tract transfer function. A block diagram of the MPLPC vocoder used in

this study is shown in Figure 3.3.1-1. In this system, the speech signal is first divided into two channels: the analysis channel, in which the LPC analysis and coding is performed; and the residual channel, in which the residual coding is performed. In the analysis channel, the first step is to apply a pre-emphasis filter of the form

$$H(z) = 1-b_1z^{-1}-b_2z^{-2} \qquad 3.3.1\text{-}1$$

where the coefficients of the filter, b_1 and b_2, have been set so as to estimate the spectral shaping effect of the glottal pulse [3.4]. The output from this filter is then used as input to an autocorrelation LPC analysis routine which performs a tenth order LPC analysis and gives an estimated vocal tract filter of the form

$$V(z) = \frac{1}{1-\sum_{n=1}^{10} a_n z^{-n}} \qquad 3.3.1\text{-}2$$

This 10th order transfer function is then both coded for transmission and, in a separate operation, corrected to include the spectral shaping effects of the pre-emphasis filter, giving the 12th order transfer function

$$V'(z) = \frac{1}{\left[1-\sum_{n=1}^{10} a_n z^{-n}\right]\left[1-b_1z^{-1}-b_2z^{-2}\right]} \qquad 3.3.1\text{-}3$$

In the residual channel, the original sampled speech signal is first passed through an all-pass filter whose transfer function is given by

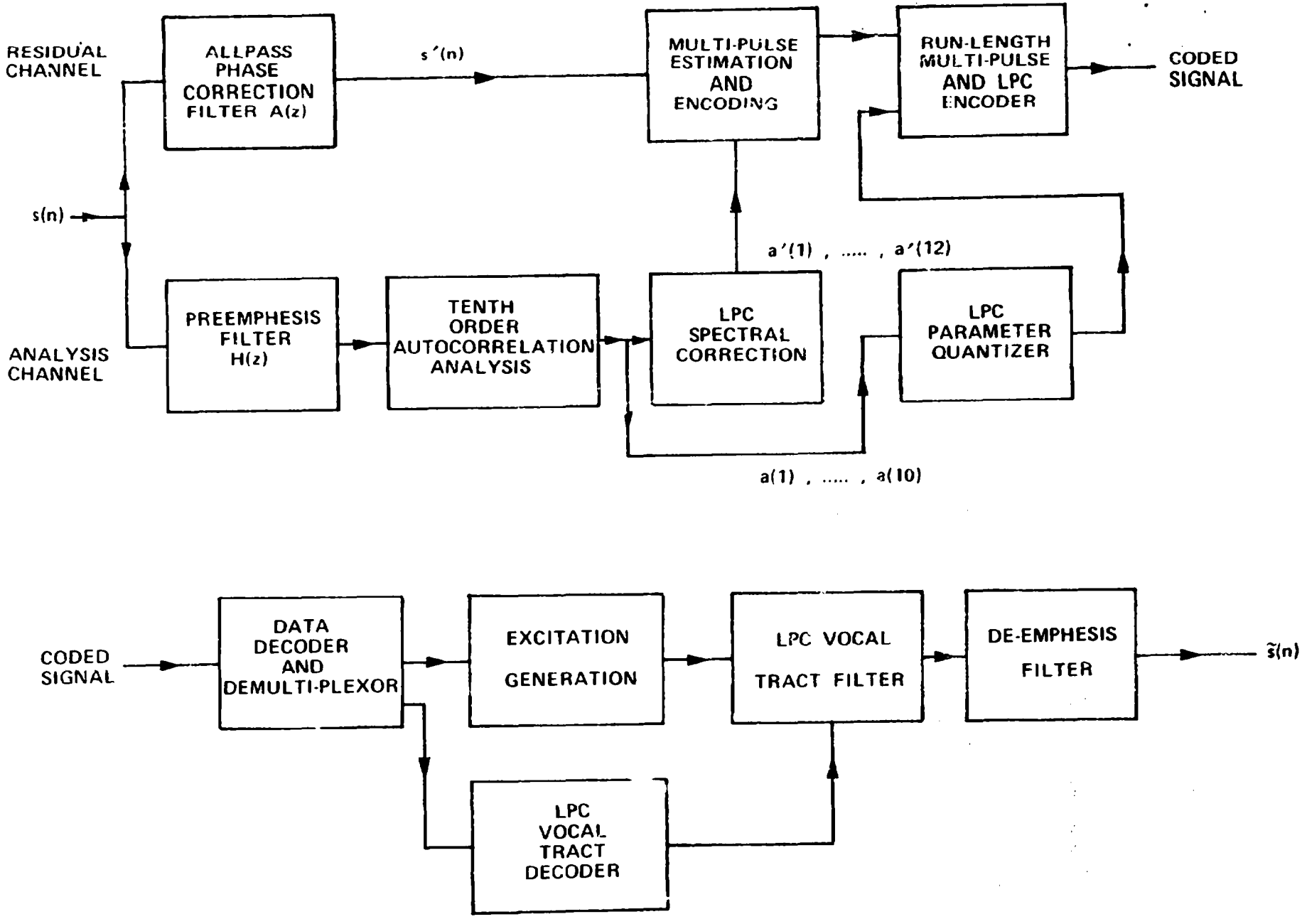


Figure 3.3.1-1. System for Generating the Multi-pulse Residual Excital Coder.

$$A(z) = \frac{b2\ z^{-2}+b\ z^{-2}+1}{1+b_1 z^{-1}+b_2 z^{-2}} \qquad 3.3.1\text{-}4$$

This filter has the effect of approximately correcting for the non-minimum phase components of the original speech signal [3.4], which in turn, has the effect of both making the speech signal more peaky in appearance and also making the vocal tract model, V'(z), more nearly correct in a phase (as well as in a spectral) sense.

The heart of the MPLPC is the Multi-Pulse Estimation and Encoding functions shown in the analysis channel in Figure 3.3.1-1. This function uses the phase corrected speech signal, s'(n), and the spectrally corrected vocal tract parameters, $a'_1 \ldots a'_{12}$, in an iterative procedure to choose a set of residual pulses to be coded and transmitted. The entire procedure is performed in frames (60 samples per frame in this study) of which only a small number of pulses are kept for transmission (2 to 10 pulses in this study). Because of the sparse nature of the multi-pulse signal, run-length coding can be used to reduce the bit rate in the MPLPC residual signal.

The iterative procedure for finding the multi-pulse locations and magnitudes used in this study can be summarized as follows. First, the ordinary residual signal $[e_0(n)]$ is formed, giving

$$e_0(n) = s'(n) - \sum_{k=1}^{12} a'(k)\ s'(n-k) \qquad 3.3.1\text{-}5$$

Next, the modified vocal tract impulse response, $h_w(n)$, is computed as

$$h_w(n)=0 \qquad n=0 \qquad \text{3.3.1-6(a)}$$

$$h_w(n) = \sum_{k=1}^{12} a'(k)\gamma^k h_w(n-k) \qquad 1< n \leq M-1 \qquad \text{3.3.1-6(b)}$$

where γ and M are control parameters of the coder. Then the modified vocal tract autocorrelation filter, $r_w(n)$, is computed as

$$r_w(n) = h_w(n) * h_w(-n) \qquad \text{3.3.1-7}$$

Using $r_w(n)$ and $h_w(n)$, the pulse locations and pulse amplitudes are computed in the following iterative procedure. First, the pulse index, p, is set to zero (p <-- 0) and $f_p(n)$ is computed as

$$f_p(n)= e_0(n) * r_w(n) \qquad \text{3.3.1-8}$$

Then the time index which maximizes $|f_p(n)|$ is found giving N_0, the location of the p^{th} pulse (for p=0 first). The approximate amplitude of the p^{th} pulse is then computed as

$$A_p = \frac{f_p(N)}{\sum_{m=0}^{M-1} h^2_w(m)} \qquad \text{3.3.1-9}$$

Once A_p is computed, the pulse index is incremented (p --> p+1), and then $f_p(n)$ is computed as

$$f_p(n) = f_{p-1}(n)-A_p r_w(n-N_p) \qquad \text{3.3.1-10}$$

The above steps are repeated until the desired number of pulse locations, $N_0 \cdots N_{P-1}$, are found. The pulse amplitudes found by this procedure are sub-optimal, and once the pulse locations are found, a new set of P amplitudes can

be found in one step [3.14].

In this study, the intent was to generate a class of distortions which were typical of MPLPC, and not specifically to implement any particular algorithm. Hence, no actual run-length coding was performed and no precise bit rates were computed. In addition, the unquantized LPC vocal tract parameters were used to generate the synthetic speech.

Another feature of the MPLPC is that once an estimate of the multi-pulse residual signal is known, it is possible to use that signal to obtain an improved estimate of the LPC vocal tract parameters. In this study, three different pulse rates (2/80, 6/80, and 10/80) were combined with original and improved LPC vocal tract parameters in order to form the six members of the MPLPC distortion sets.

3.3.2 Adaptive Transform Coder

One of the more successful methods for frequency domain speech coding is the adaptive transform coder (ATC). The basic concept on which the ATC is based involves encoding a spectral representation of the speech rather than the time domain waveform. The steps involved in the coding are: 1) windowing and transforming a segment of speech, 2) producing a model of the spectrum from LPC analysis and pitch detection, 3) dynamically allocating a predetermined number of bits among the transform coefficients using the model spectrum, and 4) adaptively quantizing the coefficients to the number of bits allocated. The decoder requires both the quantized transform coefficients and the quantized LPC parameters of the model spectrum in order to resynthesize a speech waveform. From these parameters, the bit allocations and adaptation parameters which were used in the quantizers can be computed. Resynthesis results from decoding of the transform, inverse transformation, and overlap-add combination of adjacent segments.

Our particular procedure follows closely with that of Tribolet and

Crochiere [3.18] with some modifications. The transform used in our analysis was the Discrete-Cosine-Transform (DCT) which is defined by:

$$V_c(k) = \sum_{n=0}^{M-1} v(n)c(k)\cos[(2n+1)\pi k/2M]. \qquad 3.3.2\text{-}1$$

The inverse DCT is defined as

$$v(n) = \frac{1}{M} \sum_{k=0}^{M-1} V_c(k)c(k)\cos[(2n+1)\pi k/2M], \qquad 3.3.2\text{-}2$$

where in both formulas:

$$k=0,1,\ldots,M-1 \text{ and,}$$

$$c(k) = \begin{cases} 1 & k=0 \\ 2 & k=1,2,\ldots,M-1 \end{cases} \qquad 3.5.2\text{-}3$$

Note that this transform is real, and involves computation of M equally spaced frequency components from zero to the sampling frequency. The reasons for this particular transform's use include the fact that its coefficients are always real, it is relatively simple to compute (efficient algorithms involving FFT's exist), and it is purported to be immune to windowing effects when quantized.

For the balance of the discussion, we will assume an 8kHz sampling rate for the digitized speech, since this was the case for all of the speech materials used in this study. The windows used for the analysis were 256 point trapezoids with a value of one for the center 240 points, and tapering linearly to zero on both sides. Adjacent segments were overlapped by 16 points, making an overall rate of one frame every 30 ms. A DCT of length 256 was applied for each segment.

In addition to the DCT analysis, another analysis was performed independently on data for spectrum modeling. A twelfth order LPC analysis using a 256 point windows was performed every 30 msec. Pitch detection was performed by an interactive, semi-automatic procedure to so as to minimize the probability of pitch and voicing error. These two components give rise to a smooth spectrum, $\sigma_f(k)$, and a pitch spectrum, $\sigma_p(k)$, which are combined to a model spectrum $\sigma_s(k)=\sigma_f(k)\sigma_p(k)$. The estimate $\sigma_f(k)$ was computed using a discrete Fourier transform (DFT) for the quantized linear prediction model over the first half of the unit circle. The pitch spectrum, $\sigma_p(k)$, DFT of is computed by windowing and then taking the

$$p(n) = \sum_{m=0}^{\infty} (G^{m/2})\delta(n-mL) \qquad 3.3.2\text{-}4$$

where L is the pitch and G is the ratio of the Lth lag autocorrelation term of the speech segment to the zeroth lag.

The bit assignment was a function of a weighted version of the log of $\sigma_s(k)$. This form of the bit assignment was specifically chosen so as to hide some of the quantization noise under the high energy spectral peaks. The algorithm was iterative and attempted to allocate B bits over M points, according to the formula

$$b(k) = \max\{0, \min[\mathrm{int}[\log_2(\sigma_s(k)\sigma_f^{-.25}(k)) + \delta], Nmax]\} \qquad 3.3.2\text{-}5$$

where b(k) is the number of bits assigned to transform coefficient $V_c(k)$, int[a] truncates a to an integer, and max[a,b] and min[a,b] take the maximum and minimum respectively of the two arguments, Nmax is the maximum number of bits allowed for any one coefficient, and δ is the parameter which is iteratively adjusted to make

$$\sum_{k=0}^{M-1} b(k) = B. \qquad 3.3.2\text{-}6$$

The parameters Nmax and B depend on the desired bit rate for the coding.

It is valid to assume that $V_c(k)$ is a zero mean Gaussian random variable (given only $\sigma_s(k)$ for estimation purposes) with variance equal to $\sigma_s(k)$. The quantization procedure, therefore, consists of normalizing $V_c(k)$ by $\sigma_s(k)$ and then applying a non-uniform b(k)-bit quantizer optimized for a Gaussian process of unit variance. Parameters for the quantizer were taken from Max [3.19].

In all, N bits per segment are allowed for an (Nx8000)/240=Nx33.3 bits per second rate. Of these, B bits are 'main information' and N-B bits are 'side information,' which include LPC reflection coefficients, LPC gain, pitch gain (G from equation (3.3.2-5), and pitch.

Resynthesis involves identical computation of b(k), $\sigma_s(k)$, $V_f(k)$, and $\sigma_p(k)$, which are used to calculate the quantized versions of $V_c(k)$ from the main information. An inverse DCT is then computed, and an overlap add is performed with the previous segment. The parameters use to control the adaptive transform coder are summarized in Table 3.3.2-1.

3.3.3 Subband Coder

In recent years, subband coders for digital speech coding at medium bit rates have been widely studies in the literature [3.20][3.21]. In the basic subband coding procedure (Figure 3.3.3-1), the speech is first split into frequency bands using a bank of bandpass filters. The individual bandpass signals are then decimated and encoded for transmission. At the receiver, the channel signals are decoded, interpolated, and added together to form the received signal. The subband coder derives its quality advantage by limiting the quantization noise from the encoding/decoding operation largely to the band

Bit Rate	Number Bits In Side Information Per Frame	Maximum Number of Bits Per Coefficient	Number of Bits For Transform Quantization
16 kb/s	51	5	445
12 kb/s	44	4	316
9.6 kb/s	44	4	244
8 kb/s	44	4	204
6 kb/s	44	4	136
4.8 kb/s	44	4	100

Table 3.3.2-1 Control Parameters for the Adaptive Transform Coder (ATC-2) Coding Distortion

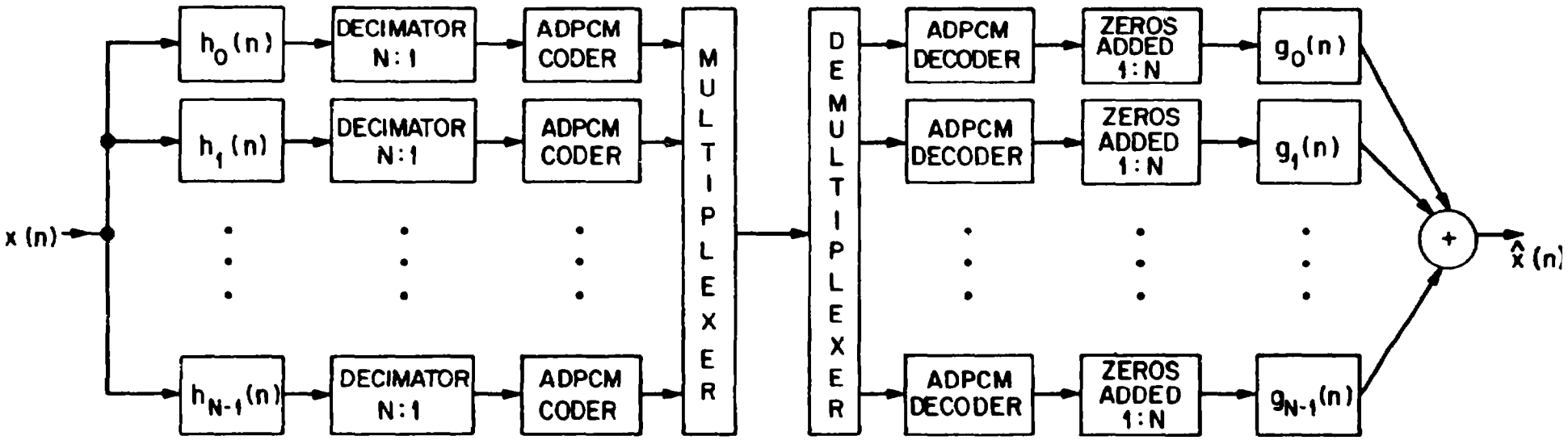


Figure 3.3.3-1. Multi-band Subband Coder

in which it is generated, thereby taking advantage of known properties of aural perception [3.22].

The basic component of octave-band tree-structured subband coders is the two-band analysis/reconstruction system shown in Figure 3.3.3-2. In this system, the analysis is performed by the two frequency selective filters, $H_0(e^{j\omega})$ and $H_1(e^{j\omega})$, which are nominally a half-band lowpass and a half-band highpass filter respectively. To preserve the system sampling rate, both channels are critically decimated at a rate of two-to-one, resulting in the two sub-sampled signals, $Y_0(e^{j\omega})$ and $Y_1(e^{j\omega})$, given by

$$Y_0(e^{j\omega}) = (1/2)[H_0(e^{j\omega/2})X(e^{j\omega/2})+H_0(-e^{j\omega/2})X(-e^{j\omega/2})] \qquad 3.3.3\text{-}1a$$

$$Y_1(e^{j\omega}) = (1/2)[H_1(e^{j\omega/2})X(e^{j\omega/2})+H_1(-e^{j\omega/2})X(-e^{j\omega/2})] \qquad 3.3.3\text{-}1b$$

In the reconstruction section, the bands are recombined, giving

$$\tilde{X}(e^{j\omega}) = (1/2)[H_0(e^{j\omega})G_0(e^{j\omega})+H_1(e^{j\omega})G_1(e^{j\omega})]X(e^{j\omega}) + (1/2)[H_0(-e^{j\omega})G_0(e^{j\omega})+H_1(-e^{j\omega})G_1(e^{j\omega})]X(-e^{j\omega}) \qquad 3.3.3\text{-}2$$

The frequency response of the two-band linear system component is contained in the first term of equation 3.3.3-2, while the second term contains the aliasing. In the classic QMF solution, the aliasing is removed by defining the reconstruction filters as

$$G_0(e^{j\omega}) = H_1(-e^{j\omega}) \qquad 3.3.3\text{-}3a$$

$$G_1(e^{j\omega}) = -H_0(-e^{j\omega}) \qquad 3.3.3\text{-}3a$$

This assignment forces the aliasing to zero, and results in a total system frequency response, $C(e^{j\omega})$, of

$$C(e^{j\omega}) = (1/2)H_0(e^{j\omega})H_1(-e^{j\omega})-(1/2)H_1(e^{j\omega})H_0(-e^{j\omega}) \qquad 3.3.3\text{-}4$$

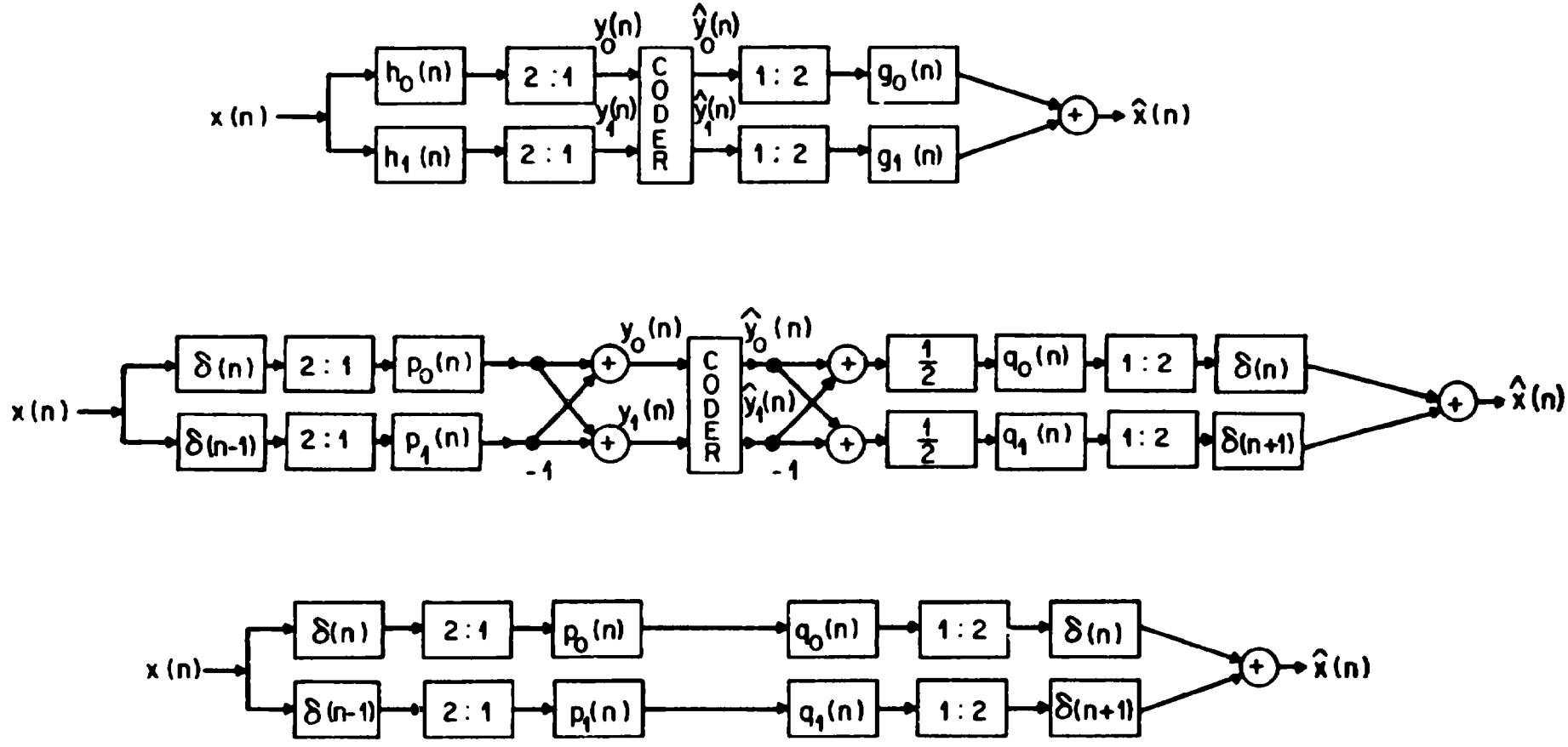


Figure 3.3.3-2. Two-Band Analysis-Reconstruction System.

In the conventional solution, the high-pass filter $[H_1(e^{j\omega})]$ and low-pass filter $[H_0(e^{j\omega})]$ are chosen to be frequency shifted versions of each other, i.e.

$$H_1(e^{j\omega})=H_0(-e^{j\omega}) \qquad 3.3.3\text{-}5$$

For this class of analysis/reconstruction system, exact reconstruction requires that

$$H_0(e^{j\omega}) - H_1(e^{j\omega}) = 2 \qquad 3.3.3\text{-}6$$

A number of authors using various methods have designed FIR filters which approximate this condition. The analysis/reconstruction systems used in this study all were based on quadrature mirror filters design by Johnston [3.23], and the systems were simulated as described by Barnwell [3.21]. The APCM coders used in this study are based on work by Jayant [3.24]. The adaptive quantizer in these systems are controlled by the dynamic steps-size $\Delta(n)$, given by

$$A(n) = \Delta(n-1) \times F[c(n-1)] \qquad 3.3.3\text{-}7$$

where $c(n)$ is the n^{th} code word and $F[\ ]$ is a preset control function. The control functions for the APCM coders used in this study are given in Table 3.3.3-1, while the control parameters for the individual systems are shown in Table 3.3.3-2.

3.3.4 Channel Vocoder

The channel vocoder which was realized was a thirty band system which occupied the frequency range of 0-3.6 kilohertz. A block diagram for each of the channels (analysis and synthesis ports) is given in Figure 3.3.4-1.

The filters in both the analysis and synthesis filter banks were all realized using recursive elliptic filters implemented as a cascade of second

APCM Coders for Subband Coding

Magnitude of Code Word [|c(n)|]

Number of Bits per Sample	0	1	2	3	4	5	6	7
4	.9	.9	.9	.9	1.2	1.6	2.0	2.4
3	.85	.9	1.4	2.0				
2	.85	1.9						

Table 3.3.3-1 Control Function F[ ] for the APCM Coders Used in the Implementation of the Subband Coders

Subband Coder Control Parameters

Coder	Number of Bands	1	2	3	4	5	Harmonic Scaling	Bit Rate
SUB-1	5	4	4	2	2	2	No	16000
SUB-2	5	3	3	2	2	2	No	14000
SUB-3	4	4	3	2	2		No	12000
SUB-4	5	4	4	2	2	2	Yes	8000
SUB-5	5	3	3	2	2	2	Yes	7000
SUB-6	4	4	3	2	2		Yes	6000

Table 3.3.3-2 Control Parameters for the Six Subband Coders Implemented as Part of This Study

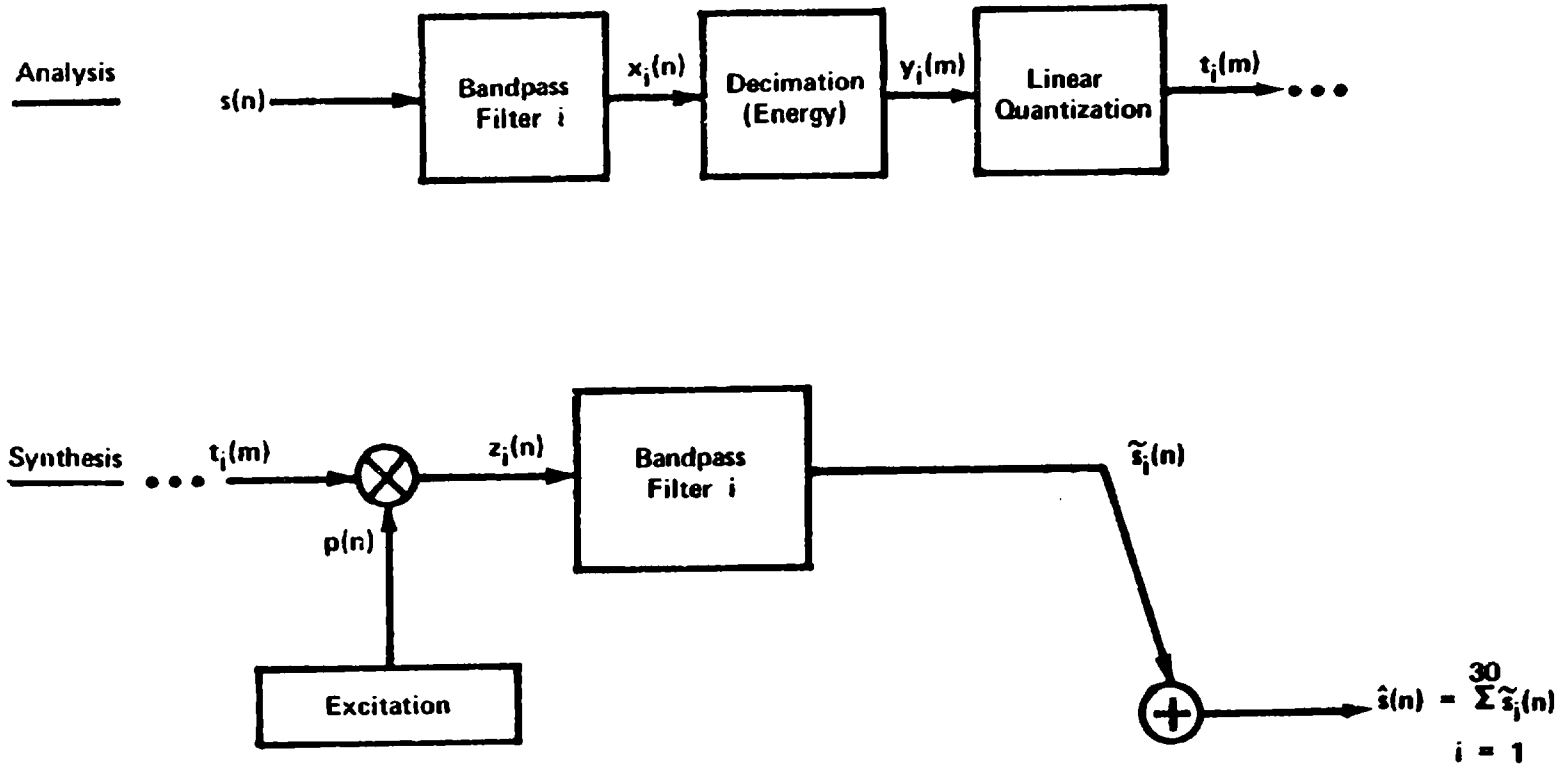


Figure 3.2.4.1 Block Diagram of Channel i.

order sections. All of the filters had an identical bandwidth of 120 Hz. The characteristics of each of the filters are given in Table 3.3.4-1. Exactly the same filters were used in the corresponding analysis and synthesis banks for each channel.

The filtered speech signal $x_i(n)$ was divided into frames of N samples. After some experimentation, N was chosen to be 215 in the final realization. Then, for each frame, the normalized square root of the energy of the windowed signal $x_i(n)$ is computed as

$$y_i(m) = \left| \frac{\sum_{n=1}^{N} [w(n)x(n)]^2}{\sum_{n=1}^{N} w^2(n)} \right|^{1/2} \qquad 3.3.4\text{-}1$$

where m is the frame number and n indexes through all the points in the frame. A Hamming window function was used used for w(n), given by

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N}\right) \qquad 3.3.4\text{-}2$$

For the channel coding, a uniform quantizer was used for the positive signal $y_i(m)$. In the final realizations, the numbers of bits used were 9,10,11,12,14 and 16 (unquantized version) respectively.

The pitch period estimations used for the channel vocoder were exactly the same as those use for the adaptive transform coder (see section 3.3.2) These pitch period signals were generated using a semi-automatic pitch detection program which minimized pitch and voicing errors. The pitch periods were estimated every 120 samples (15 msec). The excitation signal, p(n), is generated as follows: for unvoiced sounds, a uniformly distributed white random

Filter Bank for the Channel Vocoder Implementation

Filter #	Low Cutoff Frequency (kHz)	High Cutoff Frequency (kHz)	Order
1	0	.120	8
2	0.120	..240	12
3	0.240	0.360	12
4	0.360	0.480	12
5	0.480	0.600	12
6	0.600	0.720	12
7	0.720	0.840	12
8	0.840	0.960	12
9	0.960	1.080	12
10	1.080	1.200	12
11	1.200	1.320	12
12	1.320	1.440	12
13	1.440	1.560	12
14	1.560	1.680	12
15	1.680	1.800	12
16	1.800	1.920	12
17	1.920	2.040	12
18	2.060	2.160	12
19	2.160	2.280	12
20	2.280	2.600	12
21	2.600	2.520	12
22	2.520	2.640	12
23	2.640	2.760	12
24	2.760	2.880	12
25	2.880	3.000	12
26	3.000	3.120	12
27	3.120	3.240	12
28	3.240	3.360	12
29	3.360	3.480	12
30	3.480	3.600	12

Table 3.3.4-1 Filter Bank Characteristics for the Implementation of the Channel Vocoder Distortions

Control Parameters for the Channel Voocder Distortion

System Number	Bits Per Channel	Bit Rate per Channel (Bits/Second)
1	9	600
2	10	667
3	11	733
4	12	800
5	14	933
6	16	1067

Table 3.3.4-2 Control Parameters for the Channel Vocoder Distortion. For this Distortion, the Sampling Rate was 8 kHz., the Frame Size was 120 Samples, and the Number of Channels was Thirty.

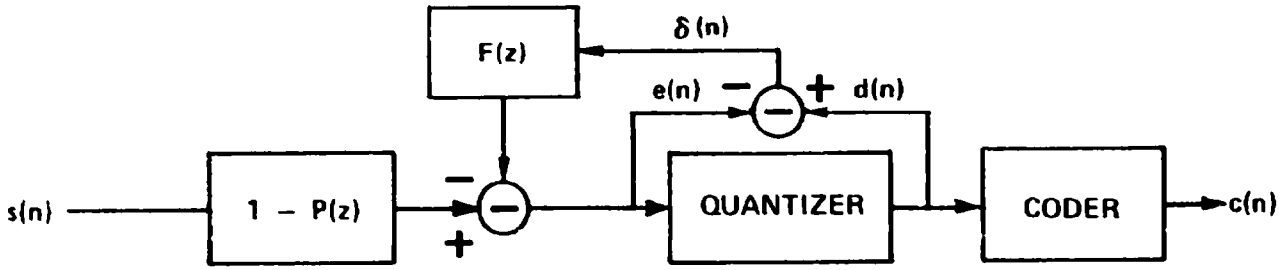
process with standard deviation G_N is used; for voices sounds, a periodic pulse train with the correct period and amplitude G_p is used. The choice of the gains G_N and G_p was critical. A ratio $G_p/G_N=10$ was found to be appropriate.

In the receiver, the excitation p(n) is multiplied by the transmitted signal $t_i(m)$ to create $z_i(n)$. This signal, in turn, is filtered to generate the channel signals, $s_i(n)$, which are all summed to create the output speech signal. The control parameters for the channel vocoder are summarized in Table 3.3.4-2

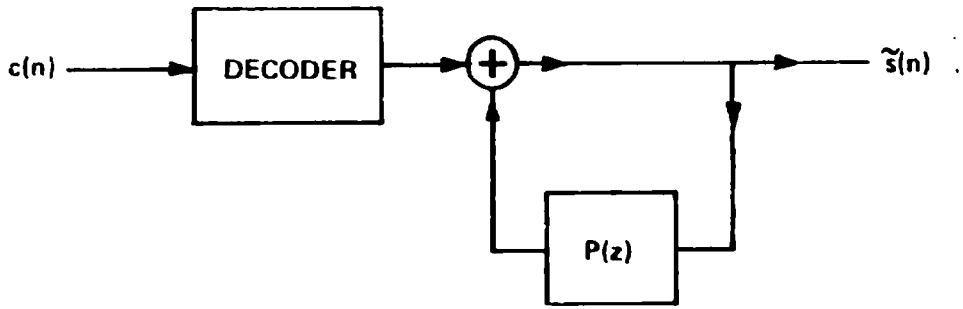
3.3.5 ADPCM with Noise Feedback

In this context, noise feedback refers to a class of analysis procedures, introduced by Atal and Schroeder [3.25], which can be applied at the transmitter of either an APC and ADPCM speech coding system in order to systematically control the spectral shape of the coding noise generated at the receiver. The reason for doing this is to take advantage of the aural noise masking effect which has been studied in psychoacoustics. This effect, compactly stated, is that in aural perception, a strong signal source will tend to mask less strong noise sources which are located close to it in frequency. Hence, it is desirable to shape the coding noise in such a way that the noise energy is placed near the speech signal energy in the short-time frequency domain.

The fundamentals of the noise feedback technique are illustrated in Figure 3.3.5-1. A key feature of this technique is that it is applied only at the transmitter of APC or ADPCM systems, and the receivers which are used are standard, unmodified APC or ADPCM receivers. Both APC and ADPCM encode a residual signal, e(n), which is obtained by passing the original signal through either a variable (APC) or fixed (ADPCM) whitening filter. In the traditional system, after quantization, the residual signal, E(z), is given by



ADPCM CODER WITH NOISE FEEDBACK



ADPCM RECEIVER

Figure 3.3.5-1. ADPCM with Noise Feedback.

$$E(z) = [1-P(z)]S(z) + [1-P(z)]\Delta(z) \qquad 3.3.5\text{-}1$$

where $P(z)$ is the transfer function of the prediction filter, $S(z)$ is the z-transform of the original speech signal and $\Delta(z)$ is the z-transform of the quantization noise signal, $\partial(n)$. At the receiver, an estimate of the original signal, $S'(z)$, is created by passing the transmitted residual signal through the inverse whitening filter, giving

$$S'(z) = E(z)/[1-P(z)] = S(z) + \Delta(z) \qquad 3.3.5\text{-}2$$

Hence, in an ordinary ADPCM or APC, the output signal is the sum of the input signal and the quantization noise signal. Since the quantization noise is nearly white, then the noise is distributed uniformly across the entire frequency band, independent of the short-time frequency spectrum of the speech.

In a noise feedback approach (Figure 3.3.5-1), the quantization noise is explicitly filtered separately from the speech signal, and the residual signal can be written as

$$E(z) = [1-P(z)]S(z) + [1-F(z)]\Delta(z) \qquad 3.3.5\text{-}3$$

giving an estimated speech signal at the receiver of

$$S'(z) = E(z)/[1-P(z)] = S(z) + \left[\frac{1-F(z)}{1-P(z)}\right]\Delta(z) \qquad 3.3.5\text{-}4$$

Hence the approximately white noise signal, $\Delta(z)$, is passed through the filter whose transfer function is given by $[1-F(z)]/[1-P(z)]$. Clearly, by varying the characteristics of $F(z)$ on a frame-by-frame basis (since $P(z)$ is always known whether it is fixed or time-varying), it is possible to shape the noise to any desired shape. An important point here is that the minimum noise energy always occurs for no noise shaping, i.e. $F(z)=P(z)$. Hence, the effect of noise

feedback is always both to shape the noise and to increase the overall noise energy.

In this study, the coding system utilized was always an ADPCM coder with a single tap fixed predictor, and the noise feedback filter was designed so that

$$\frac{1-F(z)}{1-P(z)} = \frac{1 - \sum_{n=1}^{10} a_n z^{-n}}{1 - \sum_{n=1}^{10} a_n \gamma^n z^{-n}}$$

where γ is a control parameter, and $P(z) = .9z^{-1}$. The control parameters used for this distortion are shown in Table 3.3.5-1.

3.4 Effects of Coding Distortions on Subjective Responses

3.4.1 The Effects of Multi-Pulse LPC on Subjective Responses

The effects of Multi-Pulse LPC on subjective responses are illustrated in Figure 3.4.1-1. There are several point which should be noted here. First, the Multi-Pulse LPC is capable of generating quite high quality systems at relatively low bit rates. In fact, the only coding system in this study which resulted in better quality was an ATC which operated at about twice the equivalent bit rate of the Multi-Pulse LPC. Second, the technique of using the estimated excitation function to improve the LPC analysis (systems 2, 4, and 6 of the MPLPC distortion) gives a consistent improvement for the lowest bit rates (2/80) but has little impact at the higher rates. Third, the MPLPC tends to excite a broad class of parametric distortion scales, including SF (system fluttering), SH (system highpass), SL (system lowpass), and SD (system distorted) as well as BF (background fluttering). However, on many of these scales the responses are bi-modal depending on whether there are enough pulses

Control Parameters for ADPCM with Noise Feedback

Coder	Quantizer Levels	γ	Number of LPC Taps	Predictor Coefficient
NF-1	4	.8	10	.9
NF-2	6	.8	10	.9
NF-3	8	.8	10	.9
NF-4	12	.8	10	.9
NF-5	16	.8	10	.9
NF-6	32	.8	10	.9
ADP-1	4	1	--	.9
ADP-2	6	1	--	.9
ADP-3	8	1	--	.9
ADP-4	12	1	--	.9

Table 3.3.5-1 Control Parameters for the ADPCM Systems with and without Noise Feedback Used in this Study

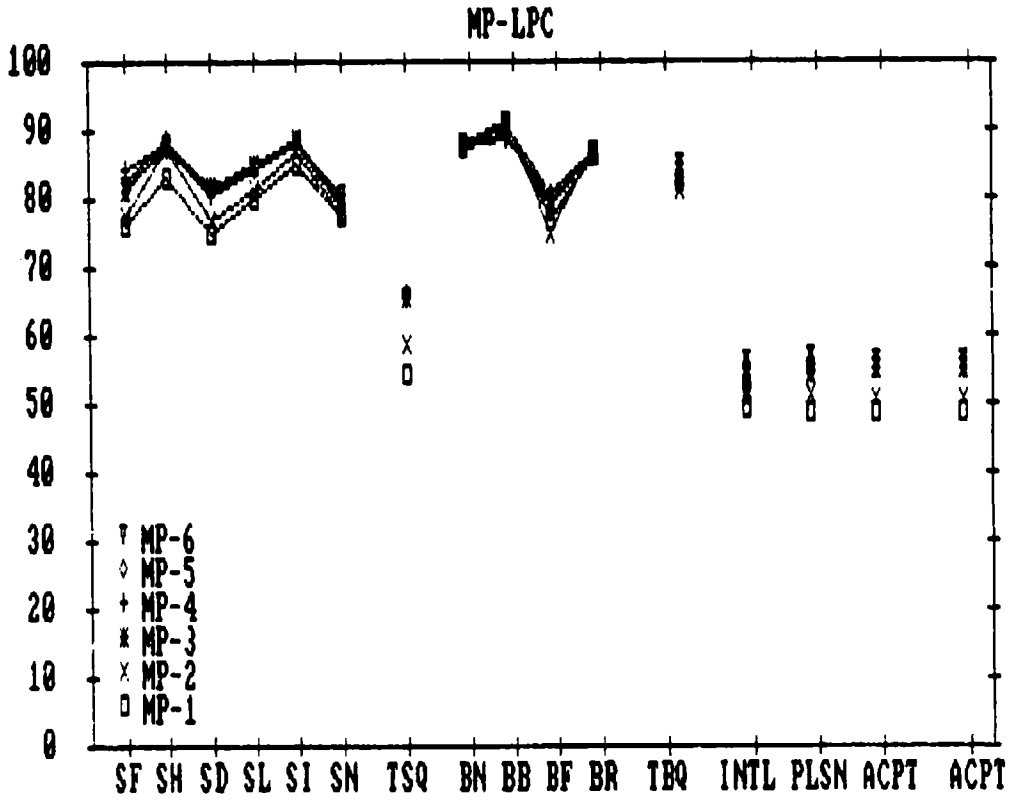


Figure 3.4.1-1 Diagnostic Acceptability Results for Multi-pulse LPC.

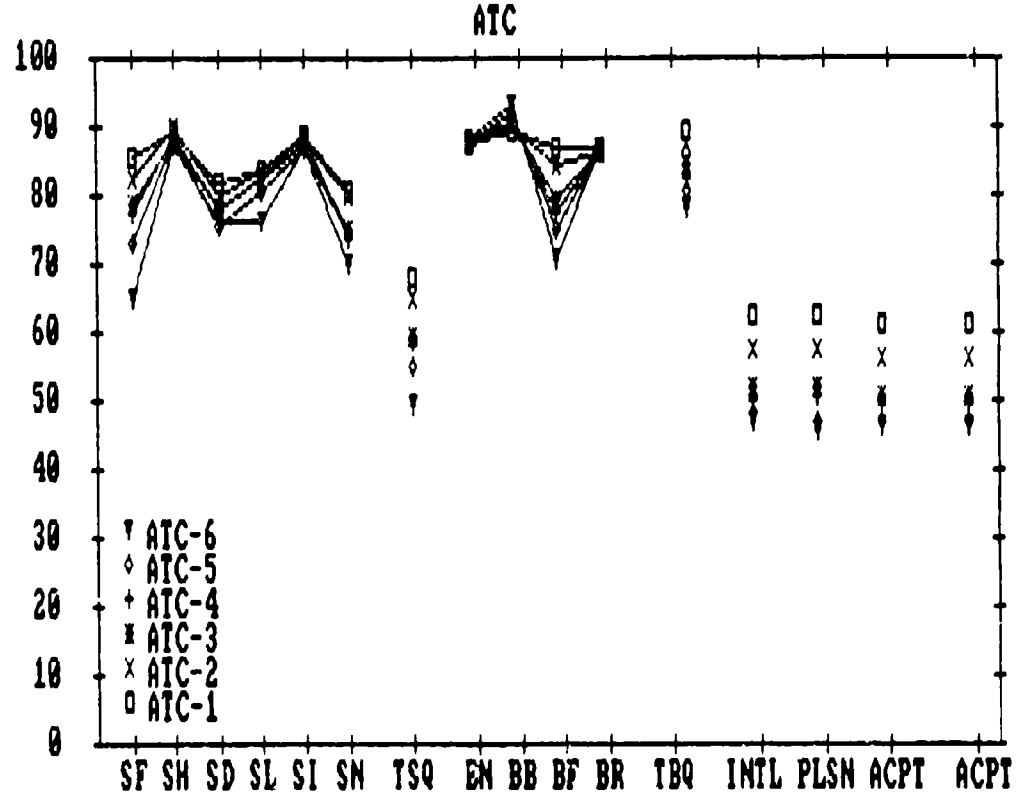


Figure 3.4.2-1 Diagnostic Acceptability Results for Adaptive Transform Coder.

in the residual representation to support the true pitch. If this effect is corrected, most of the perceived distortion occurs on the SF and BF scales.

3.4.2 The Effects of the Adaptive Transform Coder on Subjective Responses

The results of the subjective quality evaluation of the ATC is shown in Figure 3.4.2-1. The ATC clearly lives up to its billing as a high quality waveform coder for medium bit rates, with near toll quality performance at 16 Kbps. Like the MPLPC, the ATC excites a number of parametric quality scales. Clearly, the ATC distortion is mostly perceived as SF (system fluttering) and BF (background fluttering). However there are also non-trivial deviations shown on the SN (system nasal), SD (system distorted), and SL (system lowpass) scales. The spread of subjective quality results for this distortion is excellent, so the fundamental design criteria as been met.

3.4.3 The Effects of the Subband Coder on Subjective Responses

Figure 3.4.3-1 shows the results of the subband coder distortions on subjective quality. Like all of the previous distortions, the subband coder distortion exhibits a good range of subjective responses. The subband coder also exhibits a distinct bi-modal behavior for a number of parametric scales, specifically SF (system fluttering), SN (system nasal), and BF (background fluttering). This is a direct reflection of the inclusion or exclusion of time domain harmonic scaling in the subband coding system. The basic subband coder distortion shows up mostly on the SD (system distorted) scale, while the TDHS excites mostly the SF (system fluttering), SN (system nasal), and BF (background fluttering) parametric scales.

3.4.4 The Effects of the Channel Vocoder on Subjective Responses

The subjective results for the Channel Vocoder distortion are shown in Figure 3.4.4-1. Of all the coding distortions in this study, the channel vocoder was the least successful in generating a good range of subjective

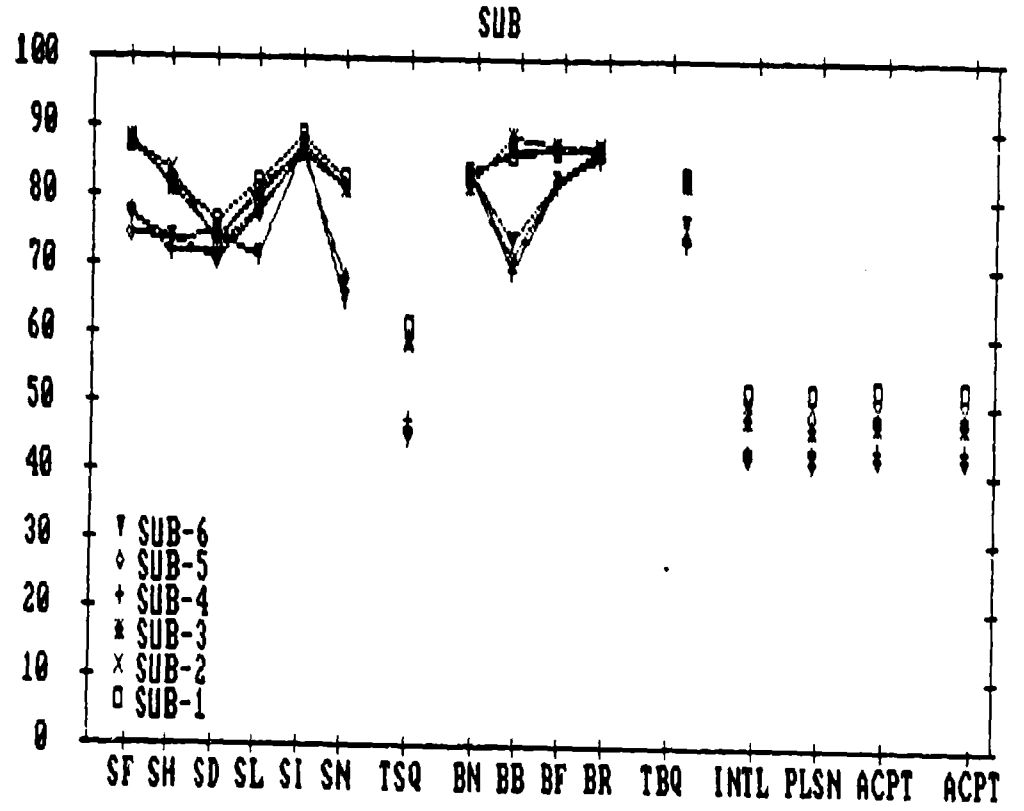


Figure 3.4.3-1 Diagnostic Acceptability Results for the Subband Coders.

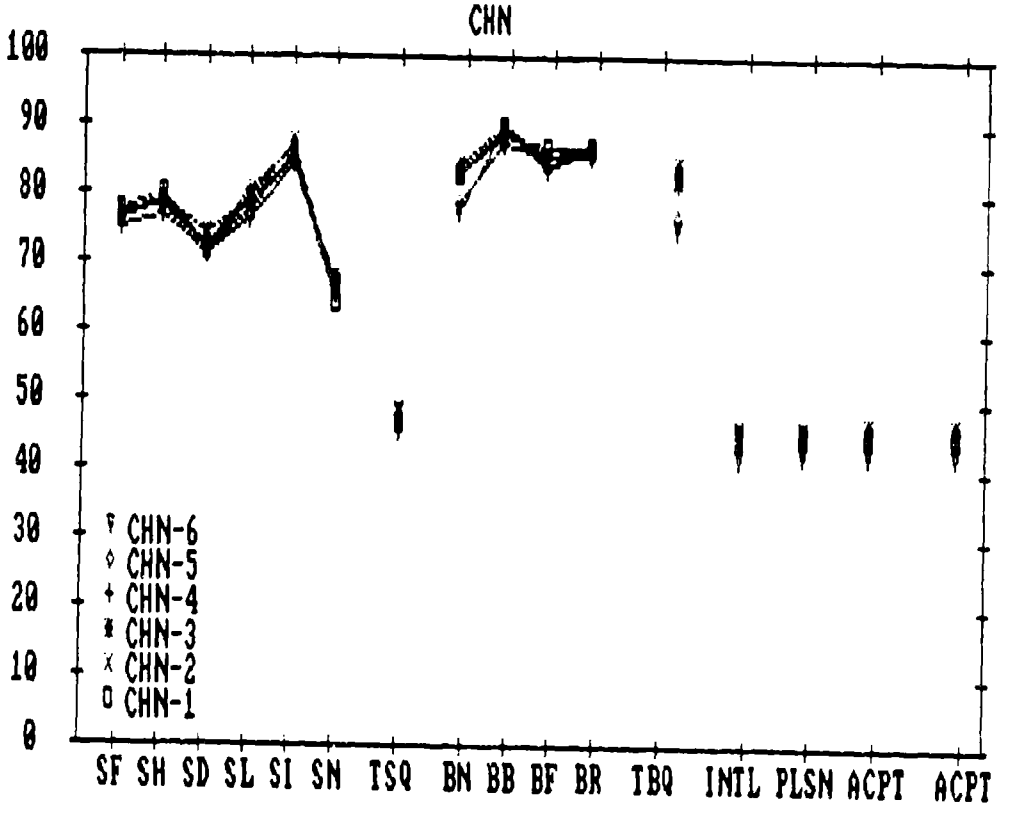


Figure 3.4.4-1 Diagnostic Acceptability Results for the Channel Vocoder.

responses. However, the results are still adequate for use in the subjective data base. It is clear from Figure 3.4.4-1 that most of the channel vocoder distortion shows up on the SN (system nasal) and BN (background noisy) scales.

3.4.5 The Effects of the ADPCM with Noise Feedback on Subjective Responses

Figure 3.4.5-1 shows the results of the subjective quality tests applied to the ADPCM-NF distortion. As can be seen from Figure 3.4.5-1, this distortion exhibits a good range of subjective responses. Almost all of the distortion shows up on the SD (system distorted) parametric scale, as is typical of many waveform coder systems. One of the claims made for the noise feedback approach is that for equivalent bit rate systems, noise feedback generally results in improved quality over systems without noise feedback. Figure 3.4.5-2 shows the results of subjective tests applied to equivalent ADPCM systems without noise feedback for the four lowest bit rate systems. Clearly, from these tests it appears that there is no measurable advantage to using noise feedback.

3.5 The Effect of the New Distortions on the Correlation Analyses

Once the new distortions were incorporated into the existing data bases, extensive tests were conducted to find the impact of the new distortions on both the correlation coefficients computed in this study and those computed in previous studies. The basic result of these analyses was that the correlation coefficients computed on the old data bases and those computed on the new data bases were very similar, and all the previously stated results were still valid for the expanded distortion ensemble.

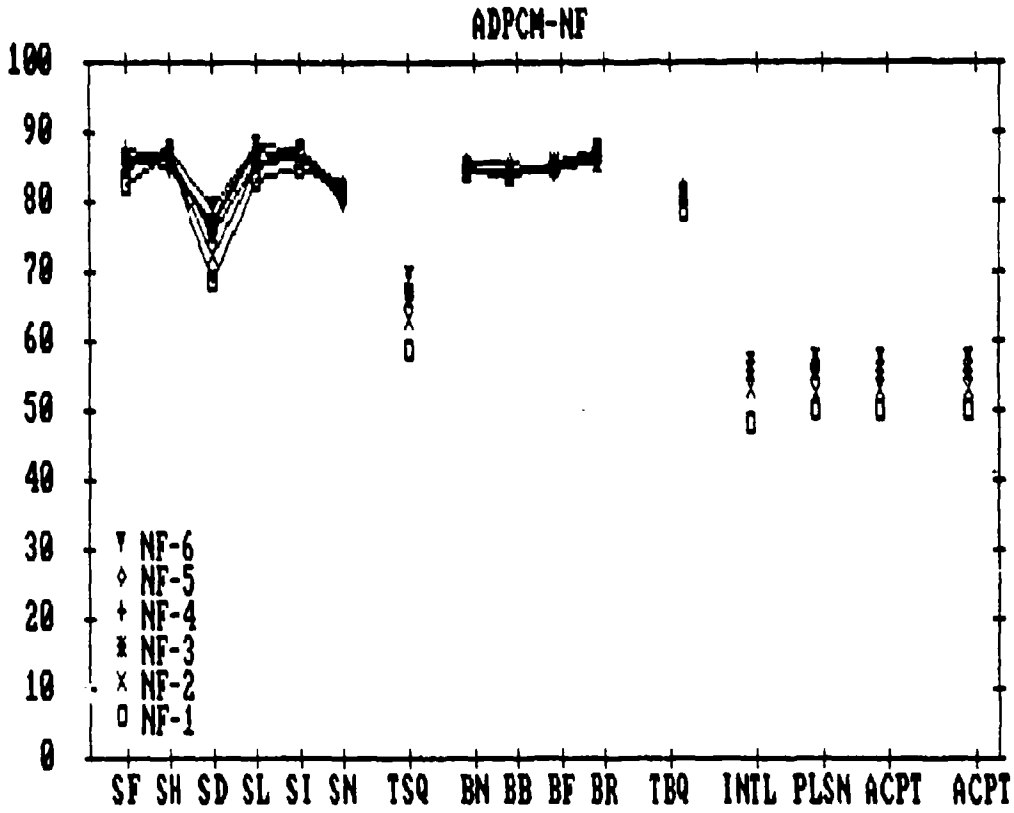


Figure 3.4.5-1 Diagnostic Acceptability Results for ADPCM with Noise Feedback.

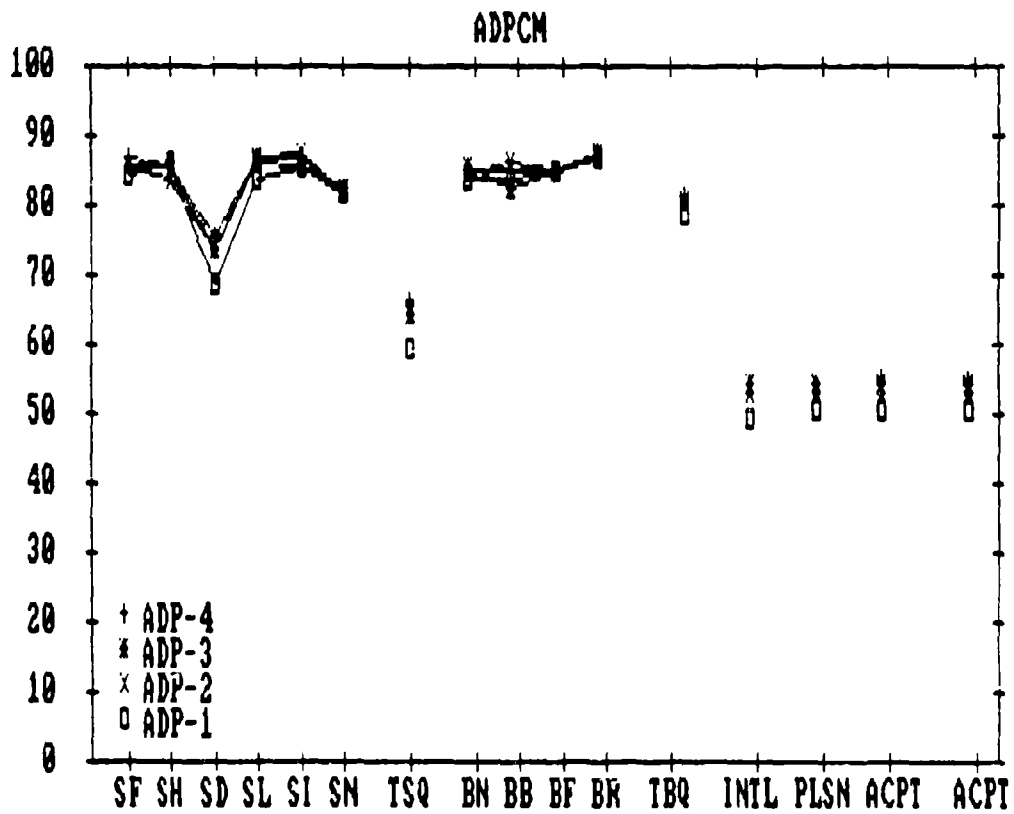


Figure 3.4.5-2 Diagnostic Acceptability Results for ADPCM without Noise Feedback.

REFERENCES

[3.1] T.P. Barnwell and W.D. Voiers, 'An Analysis of Objective Measures for User Acceptance of Voice Communications Systems,' Final Report to the Defense Communications Agency, DCA100-78-C-0003, September 1979.

[3.2] T.P. Barnwell, III, A.M. Bush, R.M. Mersereau, and R.W. Schafer, 'Speech Quality Measurement,' Final Report DCA/DCEC F30602-77-C-0118, June 1977.

[3.3] T.P. Barnwell, III, R.W. Schafer, and A.M. Bush, 'Tandem Interconnections of LPC and CVSD Digital Speech Coders,' Final Report, DCA 100-76-6-0073, 15 November 1977.

[3.4] T.P. Barnwell, III and A.M. Bush, 'Statistical Correlation Between Objective and Subjective Measures for Speech Quality,' 1978 International Conference on Acoustics, Speech, and Signal Processing, April 1978.

[3.5] T.P. Barnwell and W.D. Voiers, 'Objective Fidelity Measures for Speech Coding Systems,' presented at the meeting of the Acoustical Society of America, Honolulu, December 1978.

[3.6] T.P. Barnwell 'Objective Fidelity Measures for Speech Coding Systems,' Acoustical Society of America, Vol. 65, No. 6, December 1979.

[3.7] T.P. Barnwell, 'Correlation Analysis of Subjective and Objective Measures for Speech Quality,' 1980 International Conference on Acoustics, Speech, and Signal Processing, Denver, Colorado, April 1980.

[3.8] T.P. Barnwell, 'A Comparison of Parametrically Different Objective Speech Quality Measures Using Analysis with Subjective Quality Results,' 1980 International Conference on Acoustics, Speech, and Signal Processing, Denver, Colorado, April 1980.

[3.9] T.P. Barnwell and P. Breitkopf, 'Segmental Preclassification for Improved Objective Speech Quality Measures,' Proc. of ICASSP '81, March 1981.

[3.10] T.P. Barnwell, III, 'On the Standardization of Objective Measures for Speech Quality Testing,' Proceedings of 1982 NBS Workshop on Standards for Speech Recognition and Synthesis, Washington, DC, March 1982.

[3.11] T.P. Barnwell, III, and S.R. Quackenbush, 'An Analysis of Objectively Computable Measures for Speech Quality Testing,' Proc. of ICASSP '82, May 1982.

[3.12] S.R. Quackenbush and T.P. Barnwell, III, 'An Approach to Formulating Objective Speech Quality Measures,' Proc. 15th Southeastern Symposium on System Theory, Huntsville, Alabama, March 28-29, 1983.

[3.13] S.R. Quackenbush and T.P. Barnwell, III, 'The Estimation and Evaluation of Pointwise Nonlinearities for Improving the Performance of Objective

CHAPTER 4

MODELING OF HUMAN HEARING FOR OBJECTIVE SPEECH QUALITY ASSESSMENT

4.1 Background and Theory

Distortions of speech resulting from coding can only be detected if the magnitude of the distortion is greater than the resolution of the human auditory system. Once a distortion is perceivable, a subjective evaluation of the degree of distortion relates to the scaling properties of the auditory system. (The auditory system includes both peripheral and central components.) Our modeling approach will not deal specifically with speech perception, but rather, with the basic psychophysics of hearing. We will specifically restrict ourselves to look only at differences in coded and uncoded speech and try to quantify these differences. This approach obviously cannot address all issues, but for the coders under consideration it should be of some merit. Due to the lack of higher order modeling, it is expected that our models will more readily agree with subjective results for waveform coder type distortions than more complex distortions. Some of the key issues with hearing will be frequency, temporal, and intensity resolutions as well as their perceptual scalings.

Frequency differentiation appears to be comprised of at least two separate phenomena, one for stimuli composed of harmonically related components (pitch) and another for more general stimuli.

Pitch perception can be accurate to within 0.3%, but is applicable only to signals with specific periodicity. When the complex tones (stimuli composed of multiple sinusoids) have inharmonic components, (roughly seven or more) they cannot be perceived individually. This is the point where the pitch detection ability of human observers becomes too confused to function. Current indications are that pitch perception is a highly central neural process which must be modeled at a level much beyond the auditory periphery, and will

therefore be considered beyond the scope of our analysis.

Frequency resolution in general signals is much poorer than pitch perception for periodic signals and is determined by other basic properties. Most theories use the notion of critical bands which correspond to the presumed filtering action of the auditory system. None of the many attempts to explain psychophysical measurements of critical bands measurements solely on the basis peripheral auditory physiology up through the auditory nerve have been satisfactory. It is probable that a portion of this filtering is effected in more central neural mechanisms, and that such data as auditory nerve tuning curves would provide an incomplete model for speech perception. We therefore believe the most appropriate frequency analysis should be based on psychoacoustical measurements. Table 4.1-1 lists a set of experimentally determined critical bands which span a large fraction of the audible spectrum. Note the non-uniform bandwidths and center frequency spacing.

A well-known property of linear filters is the inverse proportionality of temporal and frequency resolution (bandwidths versus risetime). Consequently, as a filter's bandwidth increases, more precision in timing is possible. Nerve latency data suggests a lower limit for auditory resolution of around 2 ms. Low frequency stimuli give significantly worse resolution due to the corresponding narrower bandwidths of the low frequency channels, however, and temporal resolution in this range is roughly 10 ms. Although such stimuli as clicks can be resolved even when separated by as little as 2 ms, undesirable effects emerge when speech perception is modeled with such acuity. For example, pitch periods of a voiced segment of speech would be resolved. Since our analysis does not include the provision for using this information, an overall model resolution of no better than 10 ms for any channel is appropriate.

Filter Number	Center Frequency	Bandwidth
1.	50.00	70.000
2.	120.00	70.000
3.	190.00	70.000
4.	260.00	70.000
5.	330.00	70.000
6.	400.00	70.000
7.	470.00	70.000
8.	540.00	77.372
9.	617.37	86.005
10.	703.37	95.339
11.	798.71	105.411
12.	904.12	116.256
13.	1020.38	127.914
14.	1148.30	140.423
15.	1288.72	153.823
16.	1442.54	168.154
17.	1610.70	183.457
18.	1794.16	199.776
19.	1993.93	217.153
20.	2211.08	235.631
21.	2446.71	255.255
22.	2701.97	276.072
23.	2978.04	298.126
24.	3276.17	321.465
25.	3597.63	346.136

Table 4.1-1 Critical Band Center Frequencies and Bandwidths Used.

Intensity is perceived as a nonlinear function of the energies in the various critical bands. The first step of analysis is filter output envelope detection. Various mechanisms have been postulated, which include many different types of nonlinearity followed by linear filtering, resulting in a slowly varying signal for each channel. The second step involves relating the envelopes to perceived loudness, JND's (just noticeable differences), or other measures.

Masking is a mechanism undoubtedly arising from both peripheral and central processing. Critical band measurements often involve steady-state signals masking other signals, or simultaneous masking. Critical band decompositions naturally model this masking. Another form of masking occurs between signals separated in time. Most of the nonsimultaneous masking theories involve exponential decay of masking functions with time with or without frequency-dependent time constants.

4.2 Analysis Procedures

To assess the quality of coded and distorted speech using aural models, we must take into account the audibility of differences in the signals. Since we are assuming all of the distortions in the study are perceivable, the task becomes one of quantifying these differences.

The ear's frequency resolving ability strongly suggests a spectral analysis should be done to both the reference (original) speech and the distorted speech. Hence, in this study, analysis paralleling critical band filtering was performed. Of the many alternatives for the computation of the critical band-spectrum, such as LPC spectra, DFT's of windowed speech (Time dependent Fourier Transforms), and filter bank analysis, we chose the first and the last. The ear shows little sensitivity to phase as long as components are not within critical bands, and appears to respond to energy as a function of frequency. Our analysis involved short-time spectral densities. We will

denote the energy: $|V(n,s,d,m)|^2$ where n is the time index, s the speaker, d the distortion (d=ϕ means no distortion) and m is a discrete variable representing the critical band over which the energy is summed. In the LPC method, a high density DFT of the LPC spectrum is computed, and the energy in critical bands is summed. The windows for summation in the frequency domain should look like Figure 4.2-1 for auditory modeling. The pre-emphasis of roughly 3 dB/octave inherent in the wider bandwidths must be compensated. The problem with the previously mentioned computations is that although bandwidths increase with frequency, time resolution is not proportionally enhanced. To this end, we perform digital filtering and envelope detection instead, where critical band energies can be sampled faster for wider bandwidth channels than narrow ones.

Once critical band spectra were computed for original and undistorted data, comparisons were made. Sensation and auditory nerve firing rates require a nonlinear scaling of the energy envelopes. For an isolated filter's energy at an isolated time (one frame), the critical band spectral distance between the reference and distorted speech frame for that channel should be a monotonic function of the magnitude difference of the non-linearly scaled energies in the two. Here, the distance would be of the form:

$$D_m = [|f_1[V(n,s,\phi,m)] - f_1[V(n,s,d,m)]|] \qquad 4.2.1$$

where $f_1(\ )$ is a non-linearity such as a logarithm or power function. Combination of the different frequency band contributions to the overall frame distance requires both a nonlinearity applied to D_m, as well as a weighting which we assume will depend on the band's energy.

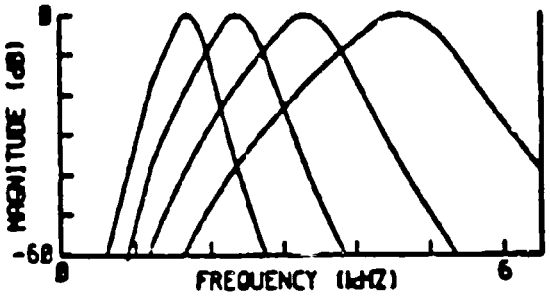


Figure 4.2-1 Critical Band Filters.

$$F_n = \sum_{n=0}^{L-1} [f_2(V(n,s,d,m) f_3(D_m)] \qquad 4.2.2$$

where the index of summation, n, covers all critical bands. Previous work suggests that $f_3(\ )$ should be $|\ |^P$ where P is a positive integer, and $f_2(\ )$ is a monotonic non-linearity such as logarithm or a power function. Combination of frames to arrive at an overall measure is accomplished in a similar manner:

$$\text{Overall distance} = \frac{\sum_n W_n f_4(F_n)}{\sum_n W_n} \qquad .3$$

where W_n is a weighting function denoting frame n's overall importance, and $f_4(\ )$ is usually the inverse function of $f_3(\ )$. In our study, we only used $f_3 = |\ |^P$ and $f_4 = |\ |^{1/P}$. Note that these choices amount to computing L_p norms for L dimensional vectors comprised of the nonlinearly transformed magnitude spectral samples.

This established framework allows for a large number of theories to be tested. The $f_1(\ )$ nonlinearity can be modeled by the JND structure for bands, or instead by the form that perceived loudness takes on as a function of intensity. In the first case, a logarithm should be used, and in the second, a non-integer power function is appropriate. By the same logic, $f_2(\ )$ should take on a similar form, although the two non-linearities need not be the same. We can also allow the functions to estimate at maximum and minimum value. As mentioned, $f_3(\ )$ and $f_4(\ )$ are of the form $|\ |^P$ and $|\ |^{1/P}$. This allows frequency based combinations to follow as L_p norm measures. Another more complex set of measures we called Klatt measures were employed, and will be

described in more detail below.

4.2.1 Log Spectral Distance Measures

According to the notion that the perceived intensity of one stimulus to another is proportional to the ratio of the two intensities (Fechner's law) or that intensity resolvability is proportional to intensity (Weber's law), the f_1 nonlinearity should be logarithmic. With the notation that F(n,s,d) is the frame distance for speaker s, frame n, and distortion d

$$F_n = F(n,s,d) = \left| \frac{\sum_{m=0}^{L-1} |V(n,s,\phi,m)|^{\gamma} \left| \log\left[\frac{V(n,s,\phi,m)}{V(n,s,d,m)}\right] \right|^{\Gamma}}{\sum_{m=0}^{L-1} |V(n,s,\phi,m)|^{\gamma}} \right|^{1/P} \qquad 4.2.1.1$$

was used.

4.2.2 Power Function Spectral Measures

Psychophysical measurements point to significant modeling errors obtained from application of Fechner's or Weber's law. A more accurate model states that the perceptual intensity doubles for every N dB increase (N is usually set to 10). Therefore, if we let i_{1p} and i_{2p} be the perceived intensities, and i_1, and i_2 be the actual intensities, the relation is:

$$\log_2\left(\frac{i_{1p}}{i_{2p}}\right) = N \log_{10}\left(\frac{i_1}{i_2}\right) \qquad 4.2.2.1$$

or

$$\frac{i_{1p}}{i_{2p}} = 2^{(N/10)\log_{10}\left(\frac{i_1}{i_2}\right)} \qquad 4.2.2.2$$

Band Number	Weight
1	.003
2	.003
3	.003
4	.007
5	.010
6	.016
7	.016
8	.017
9	.017
10	.022
11	.027
12	.028
13	.030
14	.032
15	.034
16	.035
17	.037
18	.036
19	.036
20	.033
21	.030
22	.029
23	.027
24	.026
25	.026

Table 4.2.2-1 Articulation Index Weights

$$= 2^{(N/10)(\log_2 10)\ (\log_2(\frac{i_1}{i_2}))} \qquad 4.2.2.3$$

$$= 2^{\log_2[(i_1/i_2).3N/10]} \qquad 4.2.2.4$$

$$= (i_1/i_2)^{.03N} \qquad 4.2.2.5$$

Therefore perceived intensity grows as magnitude to the .06N power. If N=10, this exponent becomes .6. A general form in which the exponent is left a free parameter, δ, would result in:

$$f_1(x) = x^{\delta} \qquad 4.2.2.6$$

Therefore:

$$F_n = \left| \frac{\sum_{m=0}^{L-1} (V(n,s,\phi,m))^{\gamma} |V(n,s,\phi,m)^{\delta} - V(n,s,d,m)^{\delta}|^P}{\sum_{m=0}^{L-1} V(n,s,\phi,m)^{\gamma}} \right|^{1/P} \qquad 4.2.2.7$$

4.2.3 Articulation Index Approximation

Although our goal is to characterize the quality of speech rather than its the intelligibility of speech, there should be some similarities in estimation methods for both. One set of procedures useful for predicting intelligibility from a description of signal to noise ratio as a function of frequency falls under the category of articulation theory. The computed value, articulation index, can be calculated in a variety of ways. Kryter's method [4.1] divides the frequency scale into 1/3-octave bands. Signal to noise ratios (SNR's in dB) are computed for each band, with a maximum of 30 dB, and a minimum of 0 dB allowed in each band. Band specific weights, listed in Table 4.2.2-1, are applied to each SNR, and these weighted values are summed.

There are a number of differences between this method and our approach. First of all, our filters are not 1/3-octave, but rather are critical bands. If anything, our analysis should be an improvement over Kryter's analysis which is only a critical band approximation. The weights which are used for the 1/3-octave filter bank can be interpolated to produce the appropriate weights for our procedure. Second, in our framework, only approximate SNR's are computed. This is accomplished by observing the differences in the original and the distorted filter bank signal energies. Third, we do not look at long term SNR's, but merely averages over many frames. With the differences kept in mind, our version of the articulation index gives a frame measure of:

$$F_n = \sum_{m=0}^{L-1} W_m \max\{0,\ \min[20\log_{10}V(\phi,m) - 20\log_{10}|V(\phi,m)-V(d,m)|, 30]\}. \qquad 4.2.3.1a$$

$$= \sum_{m=0}^{L-1} W_m N \qquad 4.2.3.1b$$

So that additional degrees of freedom could be incorporated into the model, we allowed an energy dependent frequency weighting as well as L_p norm for frequency band combinations. The resulting frame distances:

$$F_n = \left| \frac{\sum_{m=0}^{L-1} |V(n,s,\phi,m)|^{\gamma}(N)^P}{\sum_{m=1}^{L-1} |V(n,s,\phi,m)|^{\gamma}} \right|^{1/P} \qquad 4.2.3.2$$

appear similar to the log spectral distances.

4.2.4 Forward Masking Models

Simultaneous masking of signals is modeled by the critical band analysis, which describes masking as a function of frequency separation. Temporal masking, masking of one stimulus by another separated in time, also occurs. Because the effect is more dramatic when the masker precedes the target (forward masking) than the reverse (backward masking), only forward masking was considered. Various experiments indicate that masking level decays exponentially in dB with linear time [4.2] separation. The time constant for a 1000 Hz stimulus is roughly 75 ms. In other words, if the masking level of a stimulus is 80 dB at t=0, at t=75 it will be 80/e dB=30 dB. Denote τ_1 the time constant for frequency m. If the masking level at time t for frequency m and a stimulus which is no longer present is M(t,m), it would be M(t,m)/e at $t=t+\tau_1$ or $M(t+t_1,m)=M(t,m)/e$. This amounts to a frequency dependent smoothing for each filter's envelope which can be accomplished by:

$$M(n+1,s,\phi,m) = r(m)M(n,s,\phi,m)+20\log_{10} V(n+1,s,\phi,m). \qquad 4.2.4.1$$

The constant r(m) specifies the amount of smoothing and is frequency dependent. The new values, $M(n,s,\phi,m)$, can be placed into the same framework as V( ) in the log spectral distance measures.

4.2.5 Klatt-Type Measures

One interesting frame distance measure which was originally formulated for speech perception modeling has been presented by Klatt [4.3]. This measure was based on the observation that certain distortions (e.g., addition of a spectral tilt) may result in large psychoacoustic differences, but charge the perceived phonetic units very little. Four basic points were proposed by Klatt:

1) Frequency decomposition should be made which is based on critical-

bands.

2) Intensities within the frequency bands should be measured in dB SPL.

3) The slopes of the log critical-band spectra should be compared rather than the spectra themselves.

4) Differences in slopes of the log critical-band spectra should be weighted in a manner which weights peaks more than valleys.

Klatt's basic distance was of the form

$$D_{12} = \sum_{i} W(i)[S1(i)-S2(i)]^2 \qquad 4.2.5.1$$

where S1 and S2 are spectral slopes and W(i) is the weighting for each band.

By suitable adjustment of free parameters, correlation between experimentally obtained phonetic distance judgments using isolated, synthetic, steady-state vowels and the above measure achieved a correlation of .93 using this objective measure. Our feelings were that although these tasks are quite different from ours, some of the same factors may be involved in subjective phonetic distance judgments as in subjective quality evaluations.

4.3 Objective Measures

In this section we will describe the implementation of the objective measures which were introduced earlier.

4.3.1 Filter-Bank Analysis

The critical-band filters were designed in accordance with measurements and theory presented by Patterson [4.4]. Filter shapes were Gaussian, with the center frequencies and bandwidths listed in Table 4.1-1. Twenty-five filters were used to cover the spectrum 0-4000 Hz. All filters were designed using a 97-point Hamming window. Finite impulse response filtering was performed on the original and distorted waveforms, and RMS values were computed every 10

msec using a 20 msec Hamming window.

4.3.2 Frame Combinations

The main concentration of our objective measures work involved exploration of how the 10 ms frames from the distorted and original speech signals should be compared. For a given set of frame distances, F(n,s,d), objective quality was computed by simply averaging F(n,s,d) over n. In the previous study, Barnwell and Voiers had found that weighting frames by some function of their energies did not improve the performance of the objective measures tested [4.5]. We use this result as justification of our procedure.

4.3.3 Frequency Weighted Objective Measures

In the log spectral measures, frame distances were of the form shown in equation 4.2.1.1. Here L=25, and the m index denotes the different critical band channels. The free parameters were γ and P. The values used were γ=0, .2, .4, .6, 1.0, and P=.2, .5, 1, 2, 3.

The power function spectral measures were as in equation 4.2.2.7, with free parameters γ, P, and δ. The values used were γ=0, 1, ; P=1, 2, 3, ; and δ=.2, .3, .6, 1.0, 1.5, and 2.0.

The articulation index approximation as in equation 4.2.3.2 left the free parameters γ=0, .2, .4, .6, 1.0, and P=.2, .5, 1, 2, and 3. Also, in order to investigate the effect of the value of the weighting vector W listed in Table 4.2, all experiments were repeated with no weighting, i.e., a weighting vector with all elements of W equal to 1.

The forward masking models in accordance with Duifuis [4.2] allowed exponential decay of the log intensities. The frame measure was generated as shown in 4.2.1.1 but with M from equation 4.2.3.2 substituted for V. Because of earlier results, we fixed γ at 0, and let P and r(m) (specifying rate of decay - see equation 4.13) vary. The range for P was .5, 1, 2, and 3, and r(m) varied over the range 0, .2, .5, .9, .95. Note that the value 0 is the

extreme of case of no masking or a time constant of 0, and the other values lead to time constants of 6, 14, 95, and 195 ms respectively.

For the Klatt-type measures, we use Klatt's basic form as listed in equation 4.2.4.1, with slight modification. First, we define the slope of the spectrum as

$$S(n,s,d,m)=20\log_{10}[V(n,s,d,m+1)]-20\log_{10}[V(n,s,d,m)] \qquad 4.3.3.1$$

where V( ) is as before. Due to the fact that we have 25 spectral values, the index varies between 1 and 24. Not wishing to restrict ourselves to L_2 norms, we modified 4.2.4.1 to allow a free parameter, P, which gave a frame distance:

$$F(n,s,d) = \left| \sum_{m=1}^{24} W(m)\left|S(n,s,\phi,m)-S(n,s,d,m)\right|^P \right|^{1/P} \qquad 4.3.3.2$$

W( ) depends on both the distorted and original frames, and is specified by

$$W(m)=[W(\phi,m)+W(d,m)]/2, \qquad 4.3.3.3$$

where W(d,m) depends solely on the spectrum V(n,s,d,k), for k=1 to 24.

$$W(d,m) = \frac{C_1}{[C_1+\max_k V(n,s,d,k)-V(n,s,d,m)]} + \frac{C_2}{[C_2+\text{local}\max_m V(n,s,d,m)-V(n,s,d,m)]} \qquad 4.3.3.4$$

The $\max_m V(n,s,d,k)$ term indicates the maximum value V(n,s,d,k) achieves as k is varied, and $\text{local}\max_m V(n,s,d,m)$ indicates the value V(n,s,d,k) takes on at the closest peak to frequency band m. The free parameters are C_1, C_2, and P.

Values chosen were C_1=10,20,30,40,50,60,100, and 1000, C_2 = .5,1,2,10,100, and 1000, P=.5,1,2. Please note that for the cases C_1 and C_2 large, the weighting approaches 1 for all frequencies.

4.3.4 Trained Measures

Outside of critical bands, minimal auditory interaction takes place. In speech, however significant correlations exist across bands. In addition, for the set of distortions in our tests, individual frequency band distances should show some correlation with each other. A way of accounting for this would be to find the best linear combination of frequency based distances for predicting subjective quality. This procedure would amount to choosing a weighting vector, W(m),m=1,2,...,25, to maximize objective and subjective quality correlation. In this study, optimum vectors were computed for four contexts. The first two contexts weighted different frequency bands for the log spectral measure as in equation 4.2.1.1, but with the constraints γ=0 and P=2, giving the form:

$$F_n = \left| \frac{\sum_m W_m \left| \log \frac{V(n,s,\phi,m)}{V(n,s,d,m)} \right|^2}{\sum_m W_m} \right|^{1/2} \qquad 4.3.4.1$$

In one, all 25 bands were employed, whereas in the second, five bands were determined by summing filter energies in groups of 5 at a time. A similar procedure was performed for the power-law spectral distance, where γ, δ, and P for equation 4.2.2.7 were set to 0,.2, and 2 respectively, giving frame distances of:

$$F_n = \left| \frac{\sum_m W_m \left| V(n,s,\phi,m)^{\delta} - V(n,s,d,m)^{\delta} \right|^2}{\sum_m W_m} \right|^{1/2} \qquad 4.3.4.2$$

In this, both 5 and 25 band analyses were performed.

The two results of each analysis of interest are the actual weighting vector as well as the correlation achieved.

4.4 Results

The computed objective measures were calculated with the composite acceptability subjective measure described in Chapter 2. The figure-of-merit used in this portion of the research was the magnitude of the estimated correlation coefficients, ρ.

4.4.1 Log-Spectral Distance Measures

Log-spectral distance measures of the form given in equation 4.13 were tested using the free parameters given in section 4.3.3. The following observations were made.

1. For P held fixed, and γ varied, best correlation resulted from γ=0.0, for all values of P. Furthermore, the degree of correlation invariably decreased as γ moved further away from 0.0 in value.
2. For γ held fixed, and P varied, best correlation resulted from P=2 or P=3 with P=1 giving reasonably close performance. Values of P less than 1 were inferior in performance to the larger values in all cases.
3. Of the 25 combinations of parameters, the top five were:

Rank	P	γ	$\|\rho\|$ (correlation coefficient)
1.	2.0	0.0	.715
2.	2.0	0.2	.707
3.	3.0	0.0	.705
4.	1.0	0.0	.703
5.	3.0	0.2	.702

Subsets of the distortions which fit into particular categories were observed also. ADPCM and CVSD type distortions led to almost perfect correlation, as one might expect since the set is highly restrictive. Larger sets which included pole distortions, coding distortions, wide-band distortions, controlled distortions, added colored noise, added white noise, and banded distortions, were tested. Each of these included a minimum of six sets of distortions (most contained more) giving at least 144 data points for correlation analysis. Listed below are the best set of parameters for each set of distortions.

Distortion	γ	P	$\|\rho\|$
Waveform Coders (WFC)	.4	2	.71
Pole Distortions (PD)	1.	.2	.16
Coding (CODE)	0	3	.51
Wide-band (WBD)	.2	1	.58
Controlled (CON)	0	2	.72
Colored Noise (FN)	0	2	.93
Banded (BD)	0	2	.72

Most of these fit the pattern of small γ and P larger than 1. Pole distortions were not matched well at all by any set of parameters. This can be attributed to the small spread of the subjective composite acceptability

results in this set of distortions. This problem is discussed in detail in Chapter 3. In general, however, results are fairly consistent across distortions. The high correlation of objective quality with composite acceptability of added noise distortions, no doubt reflects the fact that audibility of noise and perceived quality are closely related.

4.4.2 Power Function Spectral Distance Measures

Power function distance measures with frame distances of the form given in equation 4.10 were computed with parameters listed in section 4.3.3. After running correlation analyses, the following observations were made.

1. For γ and P held fixed, correlation was always best for δ=0.2, with δ=0.3 yielding comparable but slightly worse results. In addition, as γ increased in value, performance monotonically decreased.
2. For P and δ held fixed, performance was generally best for δ=0. Only when P and δ were far from their best values did γ=1.0, give better correlation than γ=0.0, and then only slightly better.
3. For δ and γ held fixed, performance was generally superior for P=2.0, although P=1.0 and P=3.0 were not much worse.
4. The best five combinations of parameters were:

Rank	γ	δ	P	$\lvert\rho\rvert$
1.	0.0	0.2	2.0	.721
2.	0.0	0.2	1.0	.719
3.	0.0	0.3	1.0	.714
4.	0.0	0.2	3.0	.712
5.	0.0	0.3	2.0	.695

When subsets of distortions were observed as described in the previous section, the best set of parameters in terms of correlation were:

Distortion	γ	δ	P	$\lvert\rho\rvert$
WFC	0.0	0.6	1.0	.77
PD	0.0	0.6	3.0	.60
CODE	0.0	0.2	2.0	.52
WBD	0.0	0.2	1.0 or 2.0	.58
CON	0.0	0.2	2.0	.74
FN	0.0	0.2	2.0	.92
BD	0.0	0.3	1.0	.71

Again, a consistent picture emerged in that γ should be 0.0 and P could be 1.0, 2.0, or 3.0 with little difference. Only waveform coders and pole distortions led to a δ different from 0.2 or 0.3. As with log spectral measures, good prediction of colored noise distortion acceptability was possible.

4.4.3 Articulation Index

Measures of the form in equation 4.12 were tested with the parameters as described in section 4.3.2. When weighted by the vector in Table 4.2, the following results were noted.

1. Very little variation in performance existed for the entire set of parameters, with best correlation coefficients of .67 and worst .58.
2. For γ held fixed, the best value for P was either 0.2 or 0.5.
3. For P fixed, the best values of γ tended to be small, although, not always zero.
4. The top 5 systems were:

Rank	γ	P	$\lvert\rho\rvert$
1.	0.0	0.5	.67
2.	0.2	0.2	.67
3.	0.4	0.2	.67
4.	0.0	0.2	.67
5.	0.2	0.5	.67

The unweighted measures were also tested in an identical manner with the same values for the parameters. Results which were very similar to the previous tests were achieved.

1. The top 5 systems were:

Rank	γ	P	$\lvert\rho\rvert$
1.	0.2	0.2	.67
2.	0.4	0.2	.67
3.	0.0	0.2	.67
4.	0.0	0.5	.67
5.	0.6	0.2	.67

2. For P held fixed, better results where generally achieved with γ small.
3. For γ held fixed, in all cases, correlation was a monotonically decreasing function of P.
4. The spread was much larger than in the weighted case.

For the original articulation index characterization, the parameters γ=0 and P=1 should have been used. These led to scores of .65 and .64 for the weighted and unweighted cases, respectively. These values were not far from the maxima achieved. In the regular log spectral distance measure, γ=0 and P=1 led to a correlation of .70.

Distortion subsets were also tested on the unweighted measure with the following results:

Distortion	γ	P	$\lvert\rho\rvert$
WFC	0	0.2	.70
PD	1	0.2	.30
CODE	0	0.5	.63
WBD		ALL IDENTICAL	.40
CON	0	0.2	.54
FN	0	0.2	.90
BD	0	0.5	.68

For all but the pole distortions (which as mentioned earlier, gave little spread in subjective quality) small values of γ were best. The prevalence of values of P less than 1 appears throughout. For the additive colored noise distortion, as expected, good correlation was achieved.

4.4.4 Forward Masking Models

Log spectral distance measured were also formulated to use frequency dependent levels, where the levels were computed as in equation 4.1.3 with decay rates described in section 4.3.3. In all cases, for P held fixed, maximum correlation was achieved for a time constant of 0 for all channels, or no additional forward masking. The same result was observed for all the distortion subsets. The best results for the various time constants are listed below.

Time Constant	$\lvert\rho\rvert$
0 ms	.717
6	.708
14	.694
95	.675
195	.627

4.4.5 Klatt-Type Measures

Correlation tests were run on the Klatt-type measures as described in section 4.3.2. The following points were noted:

1. For all combinations of parameters C_1 and C_2, using P=1 gave superior correlation to using P=2. In most cases P=0.5 outperformed P=2, and in a few instances outperformed P=1.
2. For P fixed at 0.5, 1 and 2 rankings were as follows:

Rank	C_1	C_2	P	$\|\rho\|$
1.	10.	0.5	2	.694
2.	10.	1.0	2	.693
3.	20.	0.5	2	.691
4.	10.	2.0	2	.691
5.	30.	0.5	2	.690
6.	20.	1.0	2	.689

Rank	C_1	C_2	P	$\|\rho\|$
1.	40.	100.	1.	.736
2.	40.	1000.	1.	.736
3.	40.	10.	1.	.735
4.	50.	10.	1.	.735
5.	50.	100.	1.	.735

Rank	C_1	C_2	P	$\|\rho\|$
1.	1000.	1000.	0.5	.735
2.	100.	1000.	0.5	.734
3.	60.	1000.	0.5	.733
4.	50.	1000.	0.5	.733
5.	40.	1000.	0.5	.733

For P=0.5 or 1.0, many other combinations resulted in correlations of roughly 0.73.

The interpretation for the meaning of C_1 is that as it increases, the difference between the largest frequency band intensity and the intensity of the frequency band examined becomes less important. Similarly, as C_2 increases, the difference between the intensity of the examined band and that of the closest local maximum becomes less important. Note from equation 4.3.3.4 that since all intensities are in decibels, and differences are actually ratios, the measure is normalized for overall gain. Therefore, no terms similar to the energy weighting terms which were used in the previously described measures were used in this measure. The difference terms in equation 4.3.3.4 vary between 0 and 60, with the bulk confined to the 0 to 40 range.

The different values of P led to different choices for C_1 and C_2. In his initial experiments, for phonetic distance, Klatt essentially used only P=2. He found optimum values of C_1 and C_2 to be 20 and 1 respectively. As is evident from the table above, near maximum correlation for P=2 was achieved with just such a combination. For P=1, and C_1 fixed, C_2 tended to be larger, although a wide range was spanned. For P=1 and C_2 fixed, C_1 tended to give best results when it was roughly equal to 40. When P was 0.5, maximum correlation was achieved for C_2=1000, and C_1 large. We find it interesting that when P was 0.5, the best weighting was none at all, for P=1, the weighting was moderate, and for P=2, the best weighting was substantial. The most asthetically pleasing of these is the P=0.5, C_1=1000, C_2=1000 case, which was one of the best combinations tested. Here we see distance as a combination of square roots of differences between spectral slopes with no weighting. Differences in slopes are the same as differences between the tangents of the corresponding angles. Since the inverse tangent function has much the same

shape as the square root function, it may be that an important factor is angle, or something similar.

As with the other measures, various subsets of distortions were explored. The parameters giving best correlation for some of them are listed below:

Distortion	C_1	C_2	P	$\|\rho\|$
WFC	1000.	100.	2	.79
CODE	1000.	1000.	1	.53
WBD	1000.	0.5	0.5	.61
CON	100.	1000.	0.5	.73
FN	1000.	1000.	2	.90
BD	40.	1000.	1	.77

We observe good correlation for additive noise and waveform coder distortions. Other types of distortions were not modeled as well with a notable deficiency in coding distortions.

4.4.6 Trained Measures

Measures as described in section 4.3.4 were analyzed for optimum values for W_m. Table 4.4.6-1 lists the values achieved for the 25 and 5 band cases for log-spectral distance. Given optimum weightings, we observe substantially better performance for the 25 band case. Also, comparing optimum weighted performance with unweighted for the 25 band case, we see improvement in log-spectral measures from $|\rho|$=.72 to .78. With power-law measures, the improvement is only from .72 to .74. The five-band weighted log-spectral measure gives results close to the 25 band optimum whereas the five-band weighted power-law measure is markedly inferior.

We see no clear interpretation for the meaning of the weights in Table 4.4.6-1. The large number of zeros in the table indicates the high degree of

Band	Log Spectral Distance Weights	Power Law Distance Weights
1	-80.8	-8.4
2	106.2	7.0
3	0.0	0.0
4	0.0	19.2
5	103.1	-19.3
6	-140.5	0.0
7	0.0	-6.4
8	0.0	0.0
9	0.0	0.0
10	-32.9	2.3
11	0.0	-8.3
12	0.0	0.0
13	0.0	5.8
14	0.0	0.0
15	-27.6	-10.2
16	0.0	-2.9
17	-48.4	0.0
18	0.0	-13.3
19	0.0	33.5
20	15.5	-41.6
21	0.0	-13.5
22	-76.4	0.0
23	0.0	-17.2
24	0.0	0.0
25	25.3	12.5
Combined Band		
1	9.7	.47
2	-16.3	-1.75
3	-4.8	-1.31
4	-10.7	-1.71
5	-9.1	-1.65

Table 4.4.6-1 Trained Weights for the Trained Measures

redundancy in many of the channels for the distortion set in our data-base.

In an attempt to see if the optimum weights were robust, we conducted a few experiments. First, various subsets of distortions were evaluated for correlation of objective and subjective data. The results are listed below:

Distortion	$\|\rho\|$
WFC	.81
CODE	.60
WBD	.69
CON	.83
FN	.94
BD	.70

In almost all cases, correlations were superior to those reported in section 4.4.1. This shows that the weights give improvements pretty much across the board, giving some hope of robustness.

Another simple experiment consisted of extending the duration over which the measure was computed by roughly 40%. Objective and subjective quality were then recorrelated with a resulting coefficient of .717. This number is almost identical to that achieved with the unweighted log spectral distance over the same interval. When weighted measures were calculated over the interval not used in training, the correlation coefficient was only .56. Also unweighted log-spectral distances computed over the same interval as the weighted measures were trained on resulted in correlation of .75. The conclusion we draw from these data is that the training of the weights gives an only minor improvement (e.g., .75 to .78) when testing occurs over the same intervals used in training. When we include additional speech outside of these intervals, the trained measures lose their advantage. We feel, therefore, that the weighting

coefficients computed in training have little or no meaning in themselves.

4.5 Discussion

The measures we tested were in many cases similar to those used is previous work by Barnwell and Voiers [4.5]. The main property the auditory-based measures had in common was the critical band based spectral analysis. Various additional aspects will be examined.

Tests similar to our log-spectral and power-law measures but using uniformly spaced samples of LPC spectra were made on the same data-base by Barnwell and Voiers. In both cases, optimum parameters closely matched those observed by us. For example $\gamma=0$ in both sets of measures was best. Both studies also found the best exponent for power-law spectral distances to be 0.2. With these values the same, however, critical band spectral analyses led to correlations of .72 and .72 whereas, LPC spectral distances led to correlations of .60. Clearly the non-uniform spacing of bands was preferable. In the earlier study, non uniformly spaced LPC spectral samples were also computed by lumping 32 uniformly spaced samples into 6. Both log spectral and power-law measures achieved maximum correlations of .68, which are comparable to critical-band performance. Another factor which will be addressed shortly involves the fact that the LPC spectral analysis had poorer time resolution than the critical band analysis.

The articulation index approximation sought to measure short-time signal to noise ratios using critical band spectra. A wider class of distortions could be tested than with a time-domain short-time SNR, but at the expense of precision. This is evident from a result obtained by Barnwell and Voiers in which time domain short time SNR's had correlations of .78 with subjective acceptability of waveform coders. The articulation index measure achieved correlation of only .70 with the same subset. However, a correlation of .67 was possible for the set of all distortions where the time domain system could

only be used on a few of them. The weighting function applied to the traditional articulation index was shown in our context to give no more than slight improvement over unity weighting, which demonstrates a possible discrepancy between quality and intelligibility requirements.

The forward masking models tended to diminish the time resolution of the spectral analysis. A time constant of zero amounted to the 10 ms time resolution of the critical band analysis. Considering the degradation that occurred when this was extended to 16 ms (ρ=.717 went to ρ=.708), it may be possible that the 10 ms frames were too wide. The frequency variant measures of Barnwell mentioned above had a resolution of 15 ms. Comparing our critical band analysis smoothed to 16 ms resolution correlation result of .708 to Barnwell's .68, we see a close correspondence. In view of these facts, one may question the importance of the precision with which we formulated the spectral analysis, and argue that most any reasonable frequency variant spectral analysis choice may be virtually equivalent. The filter bank approach appears to have been worth a few percentage points in correlation, perhaps because of the increased time resolution. This could possibly be compensated for by a smaller LPC analysis windows, however.

The trained measures give an upper limit on what is possible for the particular measures tested. Although the results are hard to interpret, they allude to the fact that not all 25 filter bands are necessary. This result is highly dependent on the distortion set we used, and enough degrees of freedom existed with the weighting vector to encourage artifactual results. Again, however, this procedure tends to indicate that precise critical band analysis may be unnecessary for good results.

The Klatt-type measures performed best of all. Two factors may account for its superior results over the log-spectral measures: 1) use of spectral

slopes rather than spectral magnitude, 2) the particular weighting function used. Consider the log spectral distance with γ=0, and the Klatt measure with C_1 and C_2 large. The measures are essentially identical except for spectral slopes being used in the latter case as opposed to log spectra in the former. For P=1, the log spectral measure gives correlation of .70 and the Klatt measure gives .73. However, for P=1, the former gives p=.72 and for the latter p=.67. Therefore, simply converting from log spectra to slopes does not always lead to improvement. It should be noted, however, that given the same number of free parameters, the best Klatt-type measures outperformed the best critical band spectral distance measures. One of the best performing of the Klatt measures used unity weighting, however (with P=.5), which supports the idea that the slopes, rather than the weights, are important. Our conclusion will be that there is significant potential in this type of measure, and that it is the combination of slopes and weights which makes it unique.

4.6 Conclusion

We feel that several statements can be made in summary.

1) Simple psychophysical models do not model subjective quality extremely well. For example, the psychoacoustical growth of loudness exponent of 0.6, when put into the critical band model, performed much worse than an exponent of 0.2. Our belief is that degradations not modelable by simple distortions go much beyond the auditory periphery in their perception, and are inextricably linked to more central neural processes. The emergence of an exponent of 0.2 in several instances is quite puzzling, and possible explanations are under close scrutiny.

2) The precise Gaussian shaped critical-band filter bank characteristics may be of little importance as long as a fair number of roughly logarithmically spaced channels are used.

3) Time resolution better than 10 ms may be desirable. One suggestion is

that short windows allow differences in transient phenomena (e.g. bursts) to be measured.

4) Simple speech perception models, such as the Klatt type measures, may be of great value in the task of predicting subjective quality. Further expansion of our work to other models, we feel, has great potential.

REFERENCES

[4.1] Kryter, K.D., 'Methods for calculation and use of the articulation index,' J. Acoust. Soc. America, vol. 34, pp. 1689-1697, Nov. 1962.

[4.2] Duifuis, H., 'Consequences of peripheral frequency selectivity for nonsimultaneous masking,' J. Acoust. Soc. America, vol. 54, pp. 1471-1488, Dec. 1973.

[4.3] Klatt, D.H., 'Prediction of perceived phonetic distance from critical-band spectra: a first step,' Proceedings of International Conference on Acoustics, Speech and Signal Processing, 1982, Paris, pp. 1278-1281.

[4.4] Patterson, R.D., 'Auditory filter shapes derived with noise stimuli,' J. Acoust. Soc. America, vol. 60, pp 640-654, March 1976.

[4.5] Barnwell, T.P. and Voiers, W.D., 'An analysis of objective measures for user acceptance of voice communications systems,' Final Report, DCA Contract DA100-78-C-0003, 1979.

CHAPTER 5

PARAMETRIC OBJECTIVE MEASURES

5.1 Desirability of Estimating Subjective Parametric Quality

The purpose of any speech communications system is to permit users to communicate easily and effectively via speech. A minimum criterion for effective communication is that the speech communications link be able to reproduce a highly intelligible version of the user's speech. However, speech systems which reproduce merely intelligible speech usually do not perform well with a casual speech style, and hence are not easy to use. Higher quality speech reproduction permits a more natural speech style and promotes more effective communication since important semantic cues for speech communications, talker emotional state, or other talker qualities can be transmitted. Users can be expected to judge a speech communications system relative to their experiences in face-to-face conversation, and for each individual there will be a level of degradation for which a speech communication system will no longer be acceptable. If this minimum acceptable level is extended into a continuum of levels of acceptability, then a better criterion for easy and effective communication might be for the user to subjectively rate the system in terms of how acceptably it reproduces the user's speech.

The Diagnostic Acceptability Measure's Composite Acceptability scale is exactly this kind of subjective quality assessment (see Chapter 2). It provides valuable information for assessing quality and complexity tradeoffs in speech communication systems. Unfortunately, because of the vague and all-encompassing nature of subjective acceptability, the Diagnostic Acceptability Measure, or DAM, composite acceptability measure is difficult to track using objective measures. The quality of acceptability does not give any clues as to

the appropriate functional form for a corresponding objective measure.

There is, however, more than one quality assessment in the Diagnostic Acceptability Measure, and most of these are considerably more specific in scope then the composite acceptability scale. Table 5.1.1 lists the entire set of quality assessments which are provided by the DAM. Whereas the composite acceptability scale does not suggest a corresponding objective measure, many of the parametric subjective quality scales do. Therefore it is reasonable to expect that objective measures can be designed which will track these more specific parametric subjective qualities successfully. Once these specific objective measures are designed, they can be combined in a linear or nonlinear functional form and, using regression analysis, a measure for composite acceptability can be developed. Such objective measures would also have the advantage of providing additional diagnostic information about the nature of the perceived distortion which would not be available from an estimate of Composite Acceptability alone.

5.2 Theory

5.2.1 Multiple Linear Regression Analysis

A potentially effective procedure for combining a number of individual estimates of parametric qualities into a single estimate of Composite Acceptability is to use a multiple linear regression model. In such a model, the linear relationship between subjective and objective is hypothesized as:

$$y_i = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij} + \epsilon_i \qquad 5.2.1\text{-}1$$

where y, the dependent variable, is the isometric or parametric subjective quality and the x_j's, the independent variables, are the objective measure variables [5.1]. The β_j's are model parameters to be estimated and ϵ_i is the

DIAGNOSTIC ACCEPTABILITY MEASURE
PARAMETRIC SIGNAL QUALITIES:

INDEX	MNEMONIC	DESCRIPTORS	EXEMPLARS
SIGNAL QUALITY			
1	SF	fluttering bubbling	AM speech
2	SH	distant, thin	highpassed speech
3	SD	rasping, crackling	peak clipped speech
4	SL	muffled, smothered	lowpassed speech
5	SI	irregular, interrupted	interrupted speech
6	SN	nasal, whining	bandpassed speech
7	TSQ	total signal quality	
BACKGROUND QUALITY			
8	BN	hissing, rushing	Gaussian noise
9	BB	buzzing, humming	60 Hz hum
10	BF	chirping, bubbling	
11	BR	rumbling, thumping	low freq. noise
12	TBQ	total background quality	
TOTAL QUALITY			
13	II	raw or isometric intelligibility	
14	IP	raw or isometric pleasantness	
15	IA	raw or isometric acceptability	
16	I	parametric intelligibility	
17	P	parametric pleasantness	
18	A	parametric acceptability	
19	CA	composite acceptability	

Table 5.1-1 A list of the subjective speech quality scales in the Diagnostic Acceptability Measure.

error in the model for each observation. Subscript j is the index of the independent, or objective measure, variable and subscript i is the index of the observation, or the speaker and distortion system in the data base. Since observations in the distorted speech data base entail both a speaker and a distortion system, the observation index will more frequently indicate this explicitly as $y(s,d)$, where s indicates the speaker and d indicates the distortion system. The β_j are estimated in the classical manner by minimizing the mean square error, ϵ_i, over all distortion systems in the data base. The resulting model, which is the desired objective measure, is:

$$y = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij} \qquad 5.2.1\text{-}2$$

In order for this model to be valid, the following assumptions must be satisfied:

1. The model errors ϵ_i are uncorrelated.
2. The error ϵ has zero mean.
3. The error ϵ has constant variance σ^2.
4. The relationship between y_i and x_i is, in fact, approximately linear.

To assess the validity of these assumptions, we must investigate the source of the error term. The underlying force which determines the quality responses in the subjective data base is the types of distortions in the distorted speech data base. Therefore the distorted sentences are, fundamentally, the independent variables in that they are specified exactly. The x_i's, which are the objective measure variables, can be thought of as complex transformations of the distorted speech waveforms. Once the transformation is fixed, the x_i's are exactly specified. Therefore the error term, ϵ, should be interpreted as

error in the subjective assessment of the quality of the distorted speech samples. With this established, the above assumptions can now be evaluated.

First, the errors must be uncorrelated. In any subjective test this is insured by randomizing the order in which the data is presented for evaluation. This prevents any evaluation bias based on previous subjective judgments of similar speech segments from occurring. Dynastat Corporation used such a randomized order in the presentation of the DAM materials, so this assumption should be valid.

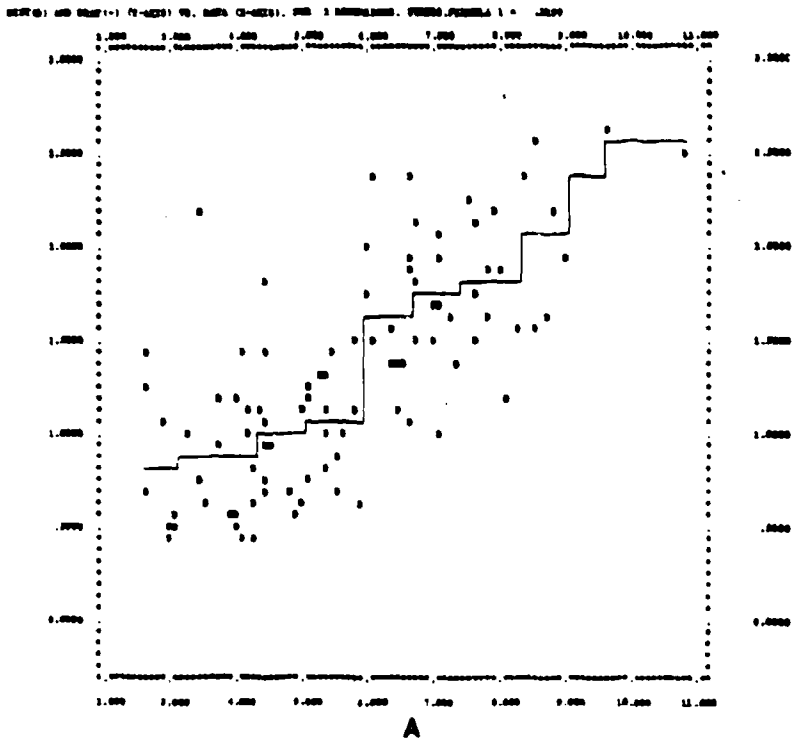
Second, the error must have zero mean, and third, the error must have constant variance. These two assumptions need to be examined together. The subjective assessments of speech quality in the subjective data base are all mean opinion scores, that is, they are based on the judgments of several individuals. Before individual opinions are averaged together, they are adjusted to eliminate the effects of that individual's preference biases (see Chapter 2). This means that each individual's assessment error is transformed to have zero mean and constant variance relative to the other listeners in the test. Furthermore, new listeners undergo a training period prior to the actual test, and can only proceed if they show a relatively small and constant quality judgment error relative to the other listeners, across a variety of distorted speech samples. Therefore, because individual judgments are adjusted to conform to a group norm and because listeners are carefully trained, assumptions two and three should be valid.

Fourth, the relationship between dependent and independent variables should be approximately linear. If this is not true, then the assumption of constant error variance will most likely be violated. In practice, one assumes that the relationship is linear, does the regression analysis, and then checks to see that the error variance is constant. This check is most easily done by

looking at a plot of error, or residual, for each observation versus the predicted value for each observation. If the model does exhibit non-constant variance, then a transformation of some or all of the x_i's may mitigate this problem. Using transformations, the relationship between independent and dependent variables can be linearized. If the residuals indicate that higher order terms in x_i are needed, these terms can be thought of as adding an additional independent variable which is simply a transformation of one of the original x_i's. In this way a polynomial model can be built within the framework of the original regression model.

5.2.2 Monotonic Regression Analysis

Monotonic regression is similar to simple linear regression in that the objective is to pass a curve through a set of points such that an objective function is minimized. In the case of monotonic regression, however, the curve need not have a parameterized functional form, such as $y = ax + b$, but rather must simply be a monotonically increasing or decreasing curve. This is a case of regression under order restrictions, and is thoroughly covered by R.E. Barlow [5.2]. In both types of regression there are three principal variables: the independent variable x_i, the dependent variable y_i, and the estimated dependent variable y_i^*, where the subscript i is the observation index. Again, in both cases the objective function to minimize is the sum of the squared error over all observations, where the error is $\epsilon_i = (y_i - y_i^*)$. However, in monotonic regression the only restriction on y_i^*, besides minimization of squared error, is monotonicity, such that $y_i^* < y_{i+1}^*$ if $x_i < x_{i+1}$. The inequality relating the y_i^*s is 'less than' for monotonically increasing regression and is 'greater than' for monotonically decreasing regression. Figure 5.2.2-1(a) shows a monotonically increasing regression curve fit and Figure 5.5.2-1(b) shows a monotonically decreasing curve fit. In these Figures x_i is the frequency index of a power spectrum, y_i. The independent variable x_i



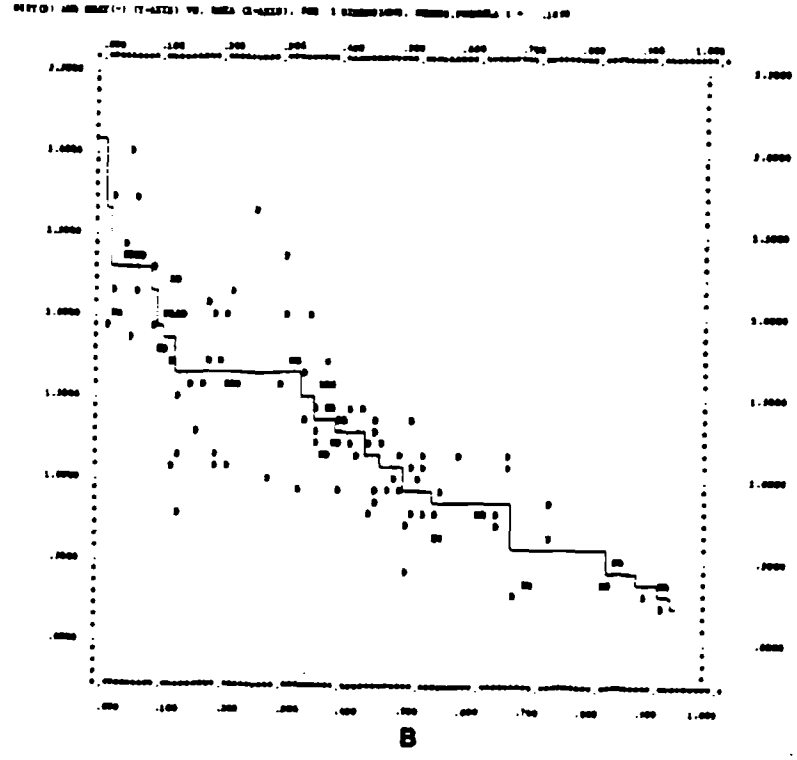


Figure 5.2.2-1 Part (a) shows a monotonicly increasing curvefit to a data set and part (b) shows a monotonicly decreasing curvefit. In both parts the x-axis is the value of $\{x_i\}$. On the y-axis the value of $\{y_i\}$ is indicated by the symbol D and the solid line is the monotonic curvefit to $\{y_i\}$.

is therefore a simple scaling of the observation index i.

The 'Up-and-Down Blocks' algorithm, developed by J.B. Kruskal [in 5.2] is an efficient method of computing a monotonic regression. Understanding the algorithm first requires defining several terms. In the following discussion assume that the dependent variable, x_i, is arranged in ascending order such that $x_i < x_{i+1}$ for all i from 1 to N-1 and that there are N elements in the dependent variable data set.

BLOCK - a set of consecutive elements y_j through y_k, $j < k$. The value of a block is equal to the average of the elements in that block.

UP-SATISFIED and DOWN-SATISFIED - consider three consecutive blocks, B-, B, and B+. For monotonically increasing regression block B is said to be up-satisfied if the average of the elements of B is less than the average of the elements in B+. For monotonicly increasing regression block B is said to be down-satisfied if the average of the elements of B is greater than the average of the elements in B-. For monotonicly decreasing regression the previous two inequalities are reversed. Additionally, any block containing y_N, is automatically up-satisfied and any block containing y_1 is similarly down-satisfied.

A flowchart of the algorithm for performing monotonic regression is shown in Figure 5.2.2-2. The algorithm begins with the independent variable data set partitioned into N blocks of one element per block. At each stage in the algorithm one block is designated as 'active'. Three choices are available for an active block. If the active block is not up-satisfied then it is combined with the next higher block. If the active block not down-satisfied then it is combined with the next lower block. If the active block is up-satisfied and down-satisfied then the next higher block becomes active. At the start the first block is active and the algorithm is terminated when the highest active block is up-satisfied. The values of the blocks at termination

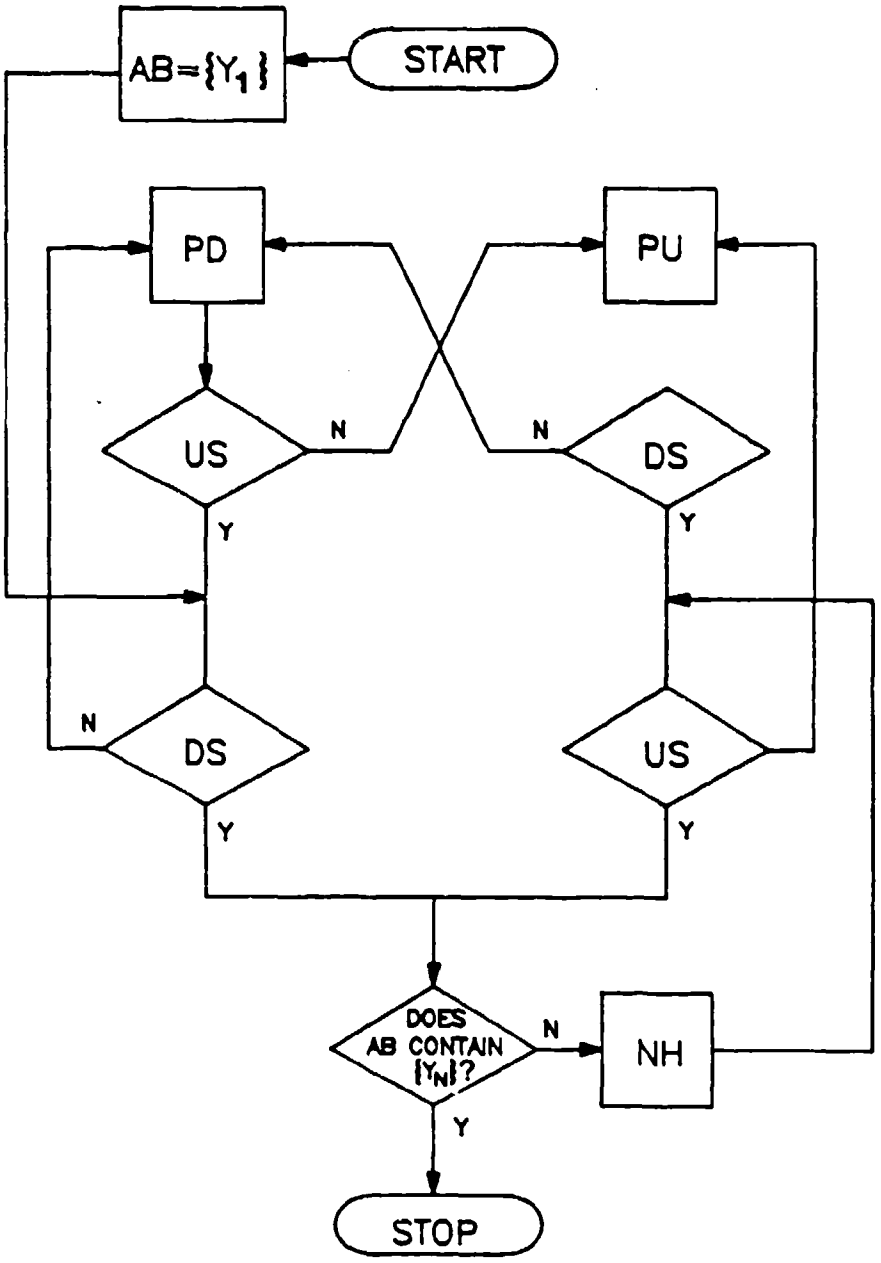


Figure 5.2.2-2 The 'Up-and-Down Blocks' algorithm. The abbreviations shown in the algorithm are described as follows:

AB - active block. At the start the block containing y_1 is the active block. The algorithm terminates if the active block contains y_N.

US - up satisfied. The conditional tests if the active block is up satisfied.

DS - down satisfied. The conditional tests if the active block is down satisfied.

PD - pool down. The current block is merged with the next lower block. This new block becomes the active block.

PU - pool up. The current block is merged with the next higher block. This new block becomes the active block.

NH - next higher. The active block is now the next higher block.

are the desired y_i^* and are the best monotonically increasing fit to the data y_i subject to minimizing the sum of the squared error. If, at termination, a block contains more that one element, for example y_j through y_k, then each corresponding estimate of the dependent variable, y_j^* through y_k^*, is equal to the value of the block containing y_j through y_k.

For the work done in this study the most significant result of monotonic regression is the statistic 'stress', which is the error variance divided by the dependent variable variance. This can be expressed as:

$$\text{Stress} = \frac{\sum_{i=1}^{N} (\epsilon_i)^2}{\sum_{i=1}^{N} (y_i - y_i)^2} \qquad 5.2.2\text{-}1$$

The stress of a monotonically increasing regression provides a measure of how closely a set of y_i's conform to a monotonically increasing function. If the set is perfectly monotonic increasing then the resultant stress is zero, and if the set is perfectly monotonic decreasing then the resultant stress is one.

An extension of monotonic regression is uni-modal regression. This regression technique fits a uni-modal curve to the data set under the constraint that the sum of the squared error is minimized. This analysis can be broken down into three steps. In the first step the mode of y_i^* is found. Assume that the observation index of the mode is M. If the mode of y_i^* is to be a global maximum, then the second step is to do a monotonically increasing regression on the points y_1 through y_M and the third step is to do a monotonically decreasing regression on the points y_{M+1} through y_N. If the mode of y_i^* is to be a global minimum, then the second step is to do a monotonically

decreasing regression on the points y_1 through y_M and the third step is to do a monotonically increasing regression on the points y_{M+1} through y_N. Stress is still expressed as in equation 5.2.2-1.

Finding the mode of y_i^* requires two monotonic regressions. As a side-note, if all intermediate results in these regressions are saved, these results being all block values for blocks 1 through the active block for each algorithm step, then the regressions required in steps two and three are already done and all three steps can be combined into one procedure. However, for the sake of clarity, the more straightforward three step approach will be described here. If the mode of y_i^* is a global maximum then a forward monotonically increasing regression and a backward monotonicly increasing regression are done. A forward regression is simply the regression performed by the up-and-down blocks algorithm. In a backward regression, however, the starting active block is y_N and the active block progresses from y_N to y_1; hence the name backward. This can be accomplished by reversing the indices on the data sets x_i and y_i, using the up-and-down blocks algorithm and then re-establishing the indices. In reversing the indices the following mapping is performed:

$$\begin{array}{lll} x_i & \text{--->} & x_{N-i+1} \\ y_i & \text{--->} & y_{N-i+1} \end{array}$$

In re-establishing the indices the same mapping is used again with the provision that the index of y_i^* is also reversed. For both forward and backward regressions the intermediate stress at each step in the algorithm must be computed. Intermediate stress values are computed using equation 5.2.2-1 with N replaced by the index of the current active block. Figure 5.2.2-3 shows the results of forward and backward regression on a data set. The curve labeled 'F' is the intermediate stress for the forward regression and the curve labeled 'B' is the intermediate stress for the backward regression. The curve labeled

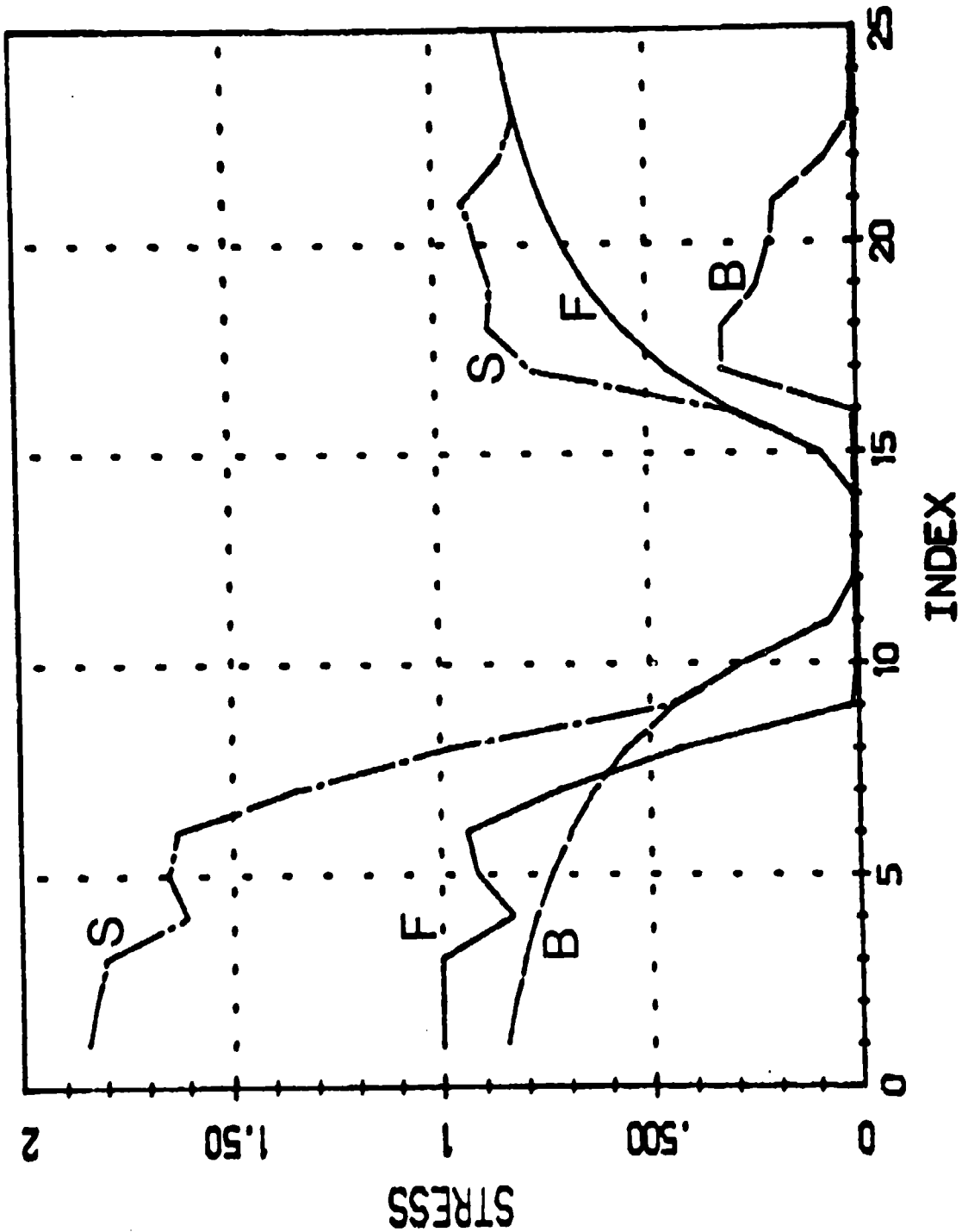


Figure 5.2.2-3 Stress curves for a unimodal maximum monotonic regression. The curve 'F' is the stress at each step for a foreward ascending monotonic regression. The curve 'B' is the stress at each step for a backward ascending monotonic regression. The curve 'S' is the sum of curves F and B at each step. The mode in the regression is the index associated with the minimum of curve S.

'S' is the sum of the two curves 'F' and 'B'. The desired quantity, the index of the mode of y_i^*, is equal to the index of the global minimum in the curve 'S', since this is the mode for which the final stress is minimum. With the mode of y_i^* established, the forward and backward regressions of steps two and three are computed as previously described and a value of stress for the unimodal regression is be computed.

5.2.3 Multidimensional Scaling

In the context of this study, multidimensional scaling, or MDS, is a tool used to graphically examine the relationship between several objective and subjective speech quality measures. It maps similarity between quality measures, as measured by correlation, into distances between quality measures as measured in an N-dimensional space. Using this technique, the relationship between many measures can be studied by examining a graph, as opposed to scanning a large table of correlation values. The principles of multidimensional scaling are best set forth by R.N. Shepard [5.3][5.4] and J.B. Kruskal [5.5][5.6]. In order to discuss the theory of multidimensional scaling, several terms need to first be defined:

OBJECT - the thing or event to be investigated. In this study objects are subjective or objective speech quality measures.

PROXIMITY - also referred to as similarity, this is a measure of the distance between objects as quantified by the magnitude of a correlation coefficient or some other distance measure.

DATA MATRIX - MDS operates on proximities associated with pairs of objects. It is convenient to think of proximities among N objects as entries in an N by N data matrix, where the entry in row i column j, m_{ij}, is the proximity of object i to object j. If we assume that the measure of proximity is a metric, then m_{ij} is equal to m_{ji} and the data matrix is symmetric. Furthermore if we assume that the proximity of an object to itself is constant for all objects,

zero for example, then the data matrix contains only (N)(N-1)/2 unique entries. For the applications in this study these assumptions are valid, so the data matrix can effectively be reduced to a lower triangular matrix of (N)(N-1)/2 proximities.

REALIZATION SPACE - the output of a multidimensional scaling is a table of coordinates which locate each object in the realization space. The distance metric in this space is Euclidean and the dimensionality of the space can be varied by the user. The distance between objects in the realization space is a function of the proximity associated with the two objects. The distance between object i and object j in the realization space is denoted as d_{ij}^*. The dimensionality of a the realization space is an important issue. For N objects it can be shown that the realization space spans at most N-1 dimensions for metric scaling and N-2 dimensions for non-metric scaling [5.7]. If the data is error free, then this dimensionality may be appropriate, though with noisy data some dimensions may be accounting for noise only. Lower dimension spaces tend to smooth out data noise since, with fewer object coordinates to estimate from the data, the coordinates have greater statistical reliability.

METRIC and NON-METRIC SCALING - scaling can be divided into these two broad categories. Mapping proximities in the data matrix into distances in the realization space in general requires a transformation on the proximities. If the function which transforms proximities to distances in the realization space is linear, then the scaling is metric. If the function is merely monotonic then the scaling is non-metric. Transformed proximities can be thought of as estimates of inter-object distances in the realization space. The transformed proximity associated with object i and object j is denoted as d_{ij}.

STRESS - points are placed in the realization space such that they minimize an error function, defined as follows:

$$\text{STRESS} = \left[\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (d_{ij} - d^{*}_{ij})}{\sum_{i=1}^{N} \sum_{j=1}^{N} (d_{ij})} \right]^{1/2} \qquad 5.2.3\text{-}1$$

STRESS measures the differences between the distance between points in the realization space and the estimated distance between points as specified by the transformed proximities. In another sense it measures how well the dimension of the realization space suits the data. The value of STRESS should guide the experimenter in choosing the appropriate dimensionality for the realization space. A rough interpretation of stress is as follows:

0%	perfect
5%	very good
10%	good
20%	fair

As an example of metric MDS, consider the data in Table 5.2.3-1 in which proximities are actually distances, in miles, between ten cities in the United States. MDS can be used with a linear transformation of the proximities (actually a simple scaling) to construct a 'map' of the U.S. as in Figure 5.2.3-1. Since this data was measured from a very nearly two dimensional space (the surface of United States land mass) the realization space need not be larger than two dimensions. In this example the STRESS, or error of fit in the realization space, would be small and nearly constant for realization space dimensionality greater than one. Figure 5.2.3-1 illustrates another important aspect of MDS: the Euclidean distance measure used in the realization space is rotation and reflection invariant, which means that the resultant configuration of points can have any angular orientation in the realization space. MDS

CITIES	ATLA.	CHIC.	DENV.	HOUS.	L.A.	MIAMI	N.Y.	S.F.	SEAT.	WASH D.C.
ATLANTA		587	1212	701	1936	604	748	2139	2182	543
CHICAGO	587		920	940	1745	1188	713	1858	1737	597
DENVER	1212	920		879	831	1726	1631	949	1021	1494
HOUSTON	701	940	879		1374	968	1420	1645	1891	1220
LOS ANGELES	1936	1745	831	1374		2339	2451	347	959	2300
MIAMI	604	1188	1726	968	2339		1092	2594	2734	923
NEW YORK	748	713	1631	1420	2451	1092		2571	2408	205
SANFRANCISCO	2139	1858	949	1645	347	2594	2571		678	2442
SEATTLE	2182	1737	1021	1891	959	2734	2408	678		2329
WASHINGTON D.C.	543	597	1494	1220	2300	923	205	2442	2329	

Table 5.2.3-1 Airline distances between ten U.S. cities [8].

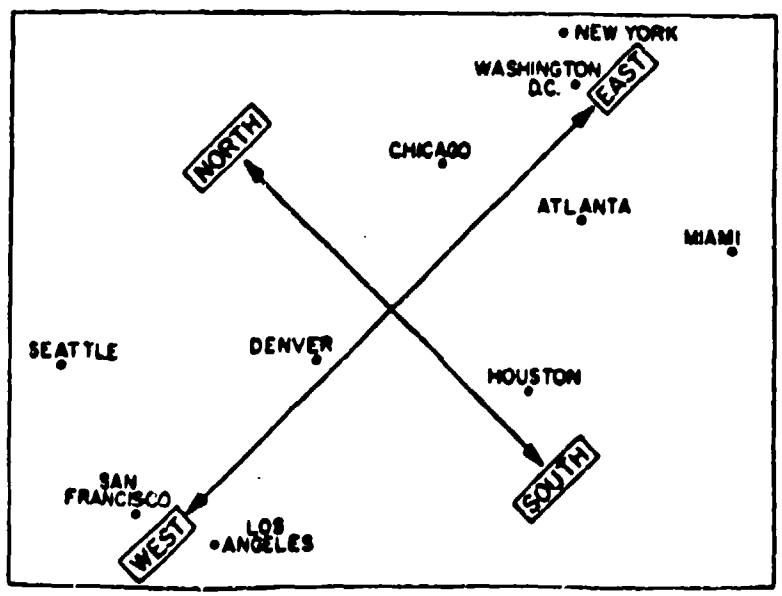


Figure 5.2.3-1 'Map' of ten cities in the U.S. as produced by multidimensional scaling of the data in table 5.2.3-1.

produces a configuration of points, but it is up to the researcher to identify the orientation and meaning of spacial dimensions in that configuration.

As an example of non-metric scaling, consider Figure 5.3.2-1(a). This is a two-dimensional scaling of the similarity between parametric quality measures in the DAM. In this scaling, parametric quality measures are the objects, and hence are represented as points in the plot. Figure 5.3.2-1(b) is a key for identification of these points. The similarity between two measures is represented by the proximity of their points in the plot. The functional measure of similarity between two measures is simply the magnitude of the correlation coefficient relating these two measures across the ensemble of distortion systems in the data base. This scaling is non-metric because the transformation of m_{ij} to yield d_{ij} is monotonic. That is, if you were to construct ordered pairs: (m_{ij}, d_{ij}), and then rank the m_{ij}'s in descending order, their corresponding d_{ij}'s would also be ranked in descending order. This is the only restriction on the transformation.

5.3 Parametric Objective Measures

5.3.1 Regression Analysis

Regression analysis has been done on the subjective quality data base by itself to determine to what extent the most desired subjective quality, composite acceptability, can be estimated from some subset of the remaining parametric subjective qualities. For two reasons only a subset of the remaining parametric qualities are considered. First, some of the subjective qualities are general in nature, rather than specific. These qualities are total signal or background quality, and overall intelligibility, pleasantness and acceptability The whole motivation for this phase of the study was to focus on narrow rather than broad quality categories, with the assumption that these would be easier to objectively estimate. Second, it is of interest, out of efficiency and expediency, to investigate how few of the parametric

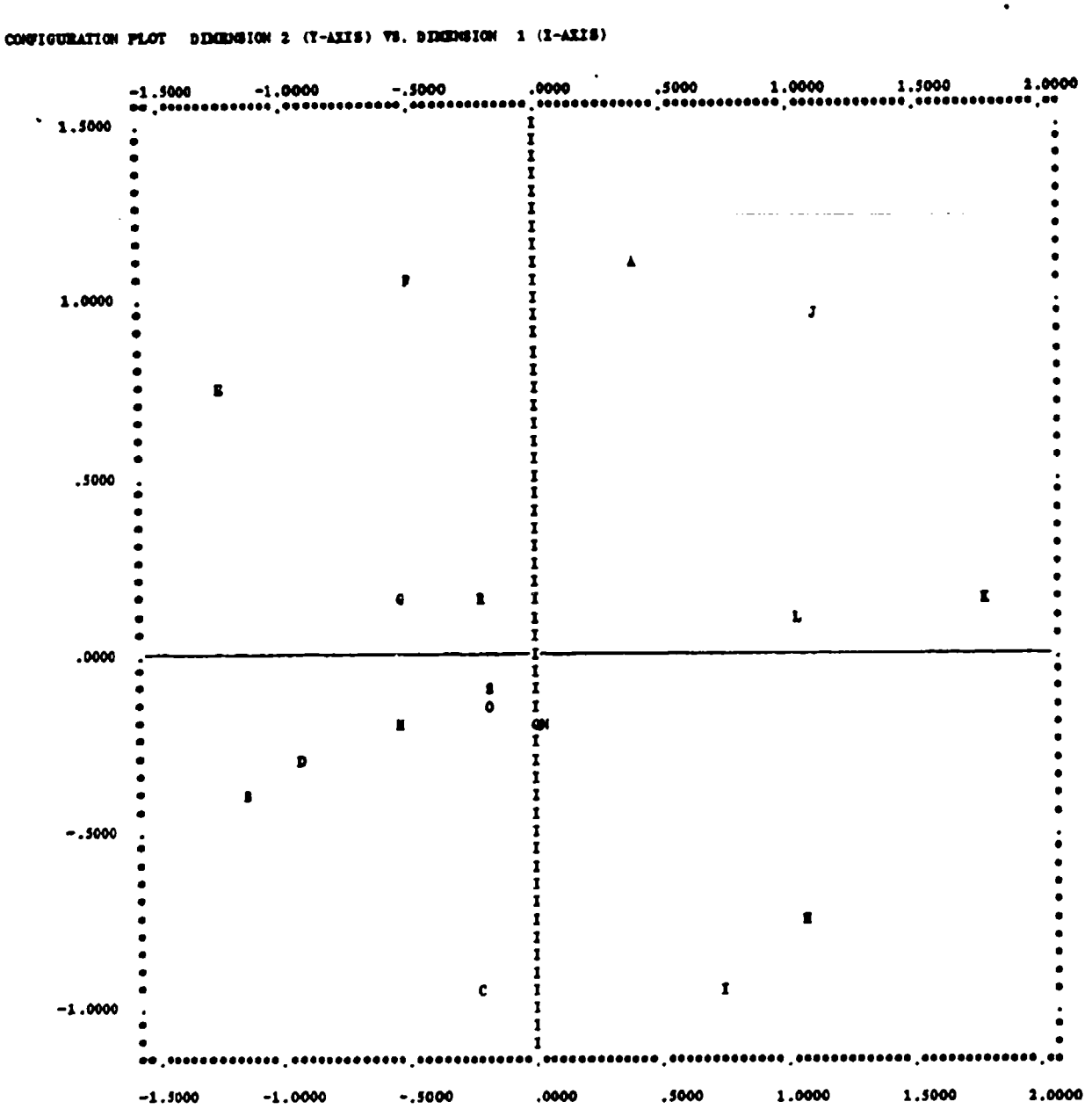


Figure 5.3.2-1(a) The results of multidimensional scaling of the subjective qualities in the Diagnostic Acceptability Measure. Points represent subjective qualities. The similarity of subjective qualities was measured by the magnitude of the correlation coefficient relating the two qualities. The plot is of a two dimensional realization space whose origin is at the center of mass of the point configuration.

SYMBOL	MNEMONIC	QUALITY
SIGNAL QUALITY		
A	SF	fluttering bubbling
B	SH	distant, thin
C	SD	rasping, crackling
D	SL	muffled, smothered
E	SI	irregular, interrupted
F	SN	nasal, whining
G	TSQ	total signal quality
BACKGROUND QUALITY		
H	BN	hissing, rushing
I	BB	buzzing, humming
J	BF	chirping, bubbling
K	BR	rumbling, thumping
L	TBQ	total background quality
TOTAL QUALITY		
M	II	raw or isometric intelligibility
N	IP	raw or isometric pleasantness
O	IA	raw or isometric acceptability
P	I	parametric intelligibility
Q	P	parametric pleasantness
R	A	parametric acceptability
S	CA	composite acceptability

Figure 5.3.2-1(b) Key to symbols in Figure 5.3.2-1(a).

subjective qualities are needed to adequately estimate composite acceptability. Fewer terms in the model for composite acceptability means fewer objective measures to build for each term and hence less computation in the composite acceptability objective measure.

The model for estimating composite acceptability from the parametric subjective qualities is identical to equation 5.2.1-2, except for these re-definition of terms: y_i is the composite acceptability score for distortion system i, x_{ij} is a parametric subjective quality score for distortion system i. In all cases the regression analysis was done over the entire 1056 distortion systems.

It should be noted that this regression analysis is simply an extraction of the model originally used by Dynastat to compute composite acceptability from the parametric subjective qualities. For this reason one should expect very good regression modeling results. This expectation was, in fact, realized by the analysis. However, good modeling results were also achieved by using only a subset of all the parametric subjective qualities to estimate composite acceptability, which is new and very encouraging information.

Three regression studies were run on the subjective data base. The first represents an upper limit on how well composite acceptability can be estimated based on all of the available information and using only linear regression models. Table 5.3.1-1(a) lists the parametric qualities used in this analysis. Note that total signal, total background, and parametric intelligibility, pleasantness and acceptability were not used because these are in fact composite qualities based on the qualities which were included in the model. The results of the analysis, listed in Table 5.3.1-1(b), is that 99.9% of the variability of composite acceptability was explained by the included variables (R-square = .9990). This is nearly perfect, indicating that the parametric subjective qualities included in the model together contain all the

INDEX	MNEMONIC	DESCRIPTORS
SIGNAL QUALITY		
1	SF	fluttering bubbling
2	SH	distant, thin
3	SD	rasping, crackling
4	SL	muffled, smothered
5	SI	irregular, interrupted
6	SN	nasal, whining
BACKGROUND QUALITY		
8	BN	hissing, rushing
9	BB	buzzing, humming
10	BF	chirping, bubbling
11	BR	rumbling, thumping
TOTAL QUALITY		
13	II	raw or isometric intelligibility
14	IP	raw or isometric pleasantness
15	IA	raw or isometric acceptability

(a)

Multiple R	.9995	Standard error of estimate	.3153
Multiple R square	.9990		

Analysis of Variance

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Regression	102252.	13	7865.	79135.
Residual	103.	1042	.0994	

(b)

Table 5.3.1-1 Part (b) shows the results of linear regression analysis with the subjective qualities listed in part (a) as independent variables and composite acceptability as dependent variable.

information present in the composite acceptability quality. As stated previously, this is to be expected since this analysis merely extracts nearly the same model used by Dynastat to compute composite acceptability.

The second analysis was limited to using only the signal and background qualities as independent variables of the regression model used to estimate composite acceptability. However this analysis was slightly different in that forward stepwise regression was used as a means of identifying the most important of these parametric qualities. As the name implies, stepwise regression is a stepwise or iterative technique used for independent variable selection. In the first step the variable which explains the most variation in the dependent variable is included in the model and all model statistics are computed. In all subsequent steps, the variable which, when added to the current model, helps explain the most variation in the dependent variable, is included in the model and all model statistics are computed. In this way, a useful, though sub-optimal, ranking of the independent variables is obtained by the degree to which the variables contribute to the model. In addition, at every step a regression model for the included independent variables is obtained.

Table 5.3.1-3 shows the results of this analysis. Listed are the parametric qualities in the order in which they entered the model, the multiple-R, or correlation coefficient, the multiple-R squared, or fraction of variability explained, and the increase in multiple-R square. The results show that two qualities dominate the rest in terms of contribution to the model. These are SD, which by itself accounts for 43 percent of the variation of CA, and SL which, along with SD, accounts for 66 percent of the variation of CA. These results are not too surprising, in that the histograms (Figures 5.4.2-1 and 5.4.2-1) for these two qualities show a much larger variance than any of the other parametric subjective qualities. Since SD and SL themselves have a

Step No.	Entered	Multiple R	Multiple R^2	Increase in R^2
1	SD, rasping, crackling	.6541	.4278	.4278
2	SL, muffled, smothered	.8120	.6594	.2316
3	SF, fluttering, bubbling	.8648	.7478	.0885
4	BN, hissing, rushing	.9039	.8171	.0692
5	P~ chirping, bubbling	.9175	.8418	.0248
6	SI, irregular, interrupted	.9380	.8798	.0380
7	SH, distant, thin	.9494	.9014	.0216
8	BB, buzzing, humming	.9518	.9059	.0045
9	BR, rumbling, thumping	.9524	.9070	.0011

Table 5.3.1-2 Results of stepwise regression. Subjective qualities are listed in the order in which they entered the model. At each step, the columns of numbers show the multiple R, multiple R-squared and increase in multiple R-squared, respectively.

large variance, they help to explain a larger portion of the variance in composite acceptability. Another encouraging result is that only seven parametric qualities are needed, SD through SH, to raise the correlation coefficient for the model above .90. Therefore only seven of the thirteen subjective qualities included in the previous regression study are needed to explain 95 percent of the variation in composite acceptability, and the remaining five subjective qualities explain less than 5 percent of the variation of composite acceptability. This analysis suggests that objective measures for only seven of the parametric subjective qualities need to be designed, since the remaining subjective qualities contribute little to the estimation of composite acceptability.

The third regression analysis was all possible subsets analysis, done to better support the conclusions reached by the stepwise regression analysis. Stepwise regression is, in general, a sub-optimal method for independent variable selection. In a given step only those variables not yet included are examined, without regard for the appropriateness of the variables already included. In contrast, all possible subsets is an optimal method of variable selection since it examines all the independent variables at each step and chooses that subset of n variables (n being the step number) which best explains the variation in the dependent variable. Therefore this analysis method will find the set of parametric subjective qualities that will yield the best estimate of composite acceptability, under the restriction that the set contain only n members.

The results of this analysis are listed in Table 5.3.1-3. For each subset of size n, the table lists the corresponding multiple R squared, multiple R and also indicates the parametric qualities included in that subset. In this method of analysis, a specific ordering of importance of parametric qualities is more difficult than with stepwise regression. Since the regression

	Parametric Quality	Number in Subset									
		1	2	3	4	5	6	7	8	9	10
1	SD, rasping, crackling	X	X	X	X	X	X	X	X	X	X
2	SL, muffled, smothered		X	X	X	X	X	X	X	X	X
4	BN, hissing, rushing				X	X	X	X	X	X	X
6	SI, irregular, interrupted					X	X	X	X	X	X
5	BF, chirping, bubbling					X	X	X	X	X	X
7	SH, distant, thin						X	X	X	X	X
3	SF, fluttering, bubbling			X	X			X	X	X	X
8	BB, buzzing, humming								X	X	X
9	BR, rumbling, thumping									X	X
10	SN, nasal, whining										X

Number in Subset	Multiple	
	R^2	R
1	0.427	0.653
2	0.659	0.812
3	0.747	0.864
4	0.816	0.903
5	0.866	0.931
6	0.885	0.941
7	0.901	0.949
8	0.905	0.951
9	0.906	0.952
10	0.906	0.952

Table 5.3.1-3 Results of all possible subsets regression analysis with the ten signal and background parametric qualities as dependent variables and composite acceptability as the independent variable. The columns of X's indicates the qualities included in the regression model for a given number of dependent variables (as indicated by the row of numbers above). For comparison, the column of numbers on the left is the order in which the parametric qualities entered the regression model in stepwise regression analysis. Below are listed the multiple R and multiple r squared for each subset of size n.

model is totally re-evaluated for each subset size, there is no one order of variable entry. The table lists parametric qualities in approximate order of entry under all possible subsets regression, and also indicates, by the numbers in the leftmost column, the order in which the qualities entered under stepwise regression. The most notable difference between the two types of analysis concerns the quality SF. Under stepwise regression this variable entered in step three, where under all possible subsets SF entered in subset three, dropped out in subset five and re-entered in subset seven. Therefore stepwise analysis overemphasizes the importance of SF. However, for the remaining parametric qualities the two analysis methods yield quite similar results.

Two conclusions can be drawn from the results of regression analysis on the subjective data base. First, that parametric subjective qualities can be used to construct a model which provides excellent estimates of subjective composite acceptability. And second, that some subset of these parametric qualities can be used to construct a model which provides estimates of composite acceptability which are nearly as good as estimates made by the full model. Given these conclusions, it is then highly desirable to construct objective measures which provide good estimates of the parametric subjective qualities, since these objective measures, combined into one large model, can be expected to provide improved estimates of subjective composite acceptability.

5.3.2 Multidimensional Scaling Analysis

Multidimensional scaling was done on the subjective data base to qualify the perceptual relationship between the parametric subjective qualities and the overall subjective qualities, and in particular composite acceptability. Figure 5.3.2-1(a) shows the results of a multidimensional scaling analysis done on the subjective data base. All nineteen subjective qualities were included in the scaling, and Figure 5.3.2-1(b) lists the key for identifying the

subjective qualities in the plot. For this analysis, the similarity between subjective qualities was equal to the magnitude of the correlation coefficient between the two qualities as computed over all the distortion systems in the data base. A descending monotonic regression was done on the similarities, so that a similarity nearly equal to 1.0 mapped into a distance nearly equal to zero. Because the transformation from similarity to distance was monotonic, the scaling was non-metric.

The analysis was done for several realization space dimensions. Figure 5.3.2-2 shows the decrease in configuration stress for increasing dimensionality. This curve does not have a distinct 'knee', where the best tradeoff between stress and dimensionality would occur, but a realization space of dimension four does yield a stress of 6 percent, which indicates a good fit. The plot in Figure 5.3.2-1(a) is for a realization space of only two dimensions, with a stress of 16.9 percent. This is rather high, indicating only a fair correspondence between the plot and the actual correlations between subjective qualities. Even so, the plot is easy to comprehend and the axes of the plot are amenable to perceptual interpretation. These two facts argue for using a two rather than four dimensional realization space, despite its high stress value.

The plot shows composite acceptability near the center of the space. The other high level qualities, intelligibility, pleasantness, and acceptability, are centered closely around composite acceptability indicating that qualities in the center of the realization space are general in nature. The left side of the realization space contains most of the signal qualities while the right side contains the background qualities, suggesting that the horizontal axis measures a signal versus background quality degradation dichotomy. Similarly, the bottom of the plot contains qualities whose exemplars are mostly fluttering or interrupted, while the top of the plot has qualities which exemplify

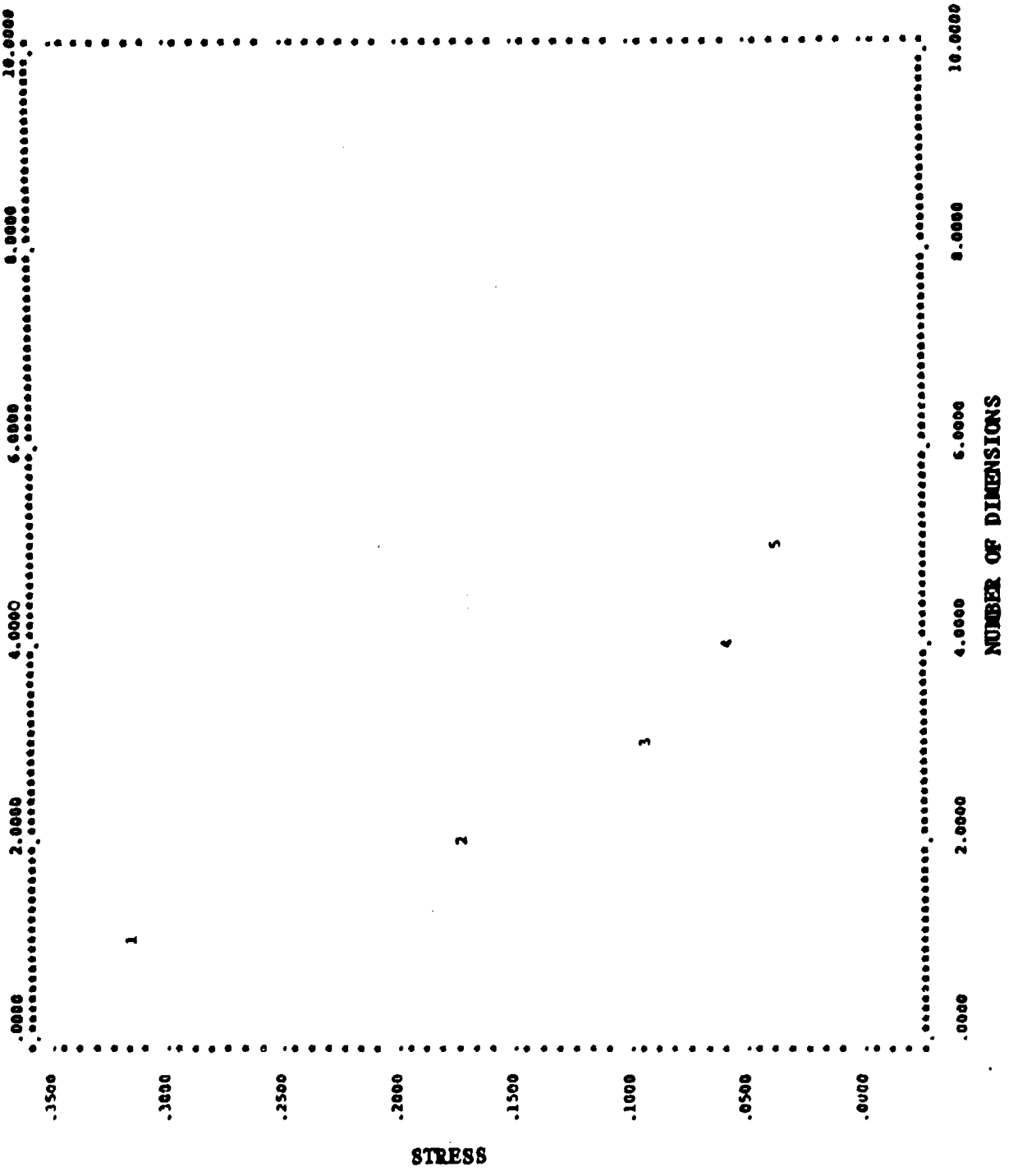


Figure 5.3.2-2 Graph of stress (y-axis) vs. dimension of realization space (x-axis) for the multidimensional scaling of figure 5.3.2-1(a).

primarily noisy distortions. Therefore the vertical axis seems to measure a noisy versus fluttering quality degradation. Finally, total signal quality and total background quality are both nearly centered within their respective signal or background parametric qualities.

One can conclude from this multidimensional scaling that the parametric qualities in the subjective data base do, in fact, measure different subjective qualities since all the parametric qualities are widely spaced in the realization space. Parametric qualities closely spaced in the realization space would indicate a large degree of redundant information. Another point is that, in two dimensions, we can associate perceptual qualities with the axes of the realization space. And finally, we note that composite acceptability is nearly in the center of the realization space, which agrees with the fact that it is an overall quality measure, and does not measure only a specific perceived quality as do those measures located near the edges of the realization space.

5.4 Parametric Objective Measures

This section of the report discusses specific objective measures which have been used to estimate parametric subjective quality. The approach used in designing an objective measure was to first understand the subjective quality it must estimate. The subjective scores provide a key to this understanding. Distortions which register a subjective quality score widely deviating from the average are exemplary of that quality, and hence provide insight into the physical or objective nature of that subjective measure. This approach to understanding the meaning of subjective quality scales will be discussed in detail for each of the parametric qualities identified as most important by the regression analysis in section 5.3.1. Before proceeding, however, the meaning of the term 'distortion' should be clarified. In the distorted speech data base, each distortion is comprised of four talkers with six distortion levels for each talker (Chapter 2). In the following analysis these 24 distortion

systems are grouped together and are referred to simply as a distortion.

5.4.1 SD: Rasping, Crackling

This subjective quality describes the degree to which speech is rasping or crackling. Table 5.4.1-1 lists the distortions which excite the system distorted scale. For each distortion the minimum, maximum and range of quality scores associated with that distortion are listed. The degree to which a distortion exemplifies a parametric quality is related to either the range, or spread, of the distortion on the parametric quality scale or to the maximum quality score on that scale. The latter case occurs when a distortion does not have a large range, but instead scores uniformly low on the subjective quality scale, and therefore indicatives that the entire distortion exemplifies that quality. The list in Table 5.4.1-1 is ordered according to the range of the distortion quality scores so that, in general, the distortions most exhibiting the subjective quality fall at the top of the list.

The dominant physical or objective characteristics the distortions in Table 5.4.1-1 have in common is that they involve nonlinearities which distort the waveform and therefore smear energy across the spectrum. This smearing is particularly noticeable at higher frequencies where the speech level is naturally lower and more easily dominated by noise from nonlinearities. Also present are additive noise distortions, which bolster the hypothesis that noise, either correlated to the speech power and arising from nonlinearities or uncorrelated and arising from an additive process, is the objective character of this subjective quality.

As mentioned in section 5.3.1, system distorted accounts for a very large fraction of the variance of composite acceptability, some 60%. This is principally because of the large number of distortions which excite the system distorted scale. The histogram in Figure 5.4.1.1 gives another perspective on

SD rasping, crackling

DISTORTION	MAX	MIN	RANGE
center clipping	83.90	50.70	33.20
CVSD	85.40	53.40	32.00
ADM	85.40	57.70	27.70
peak clipping	81.50	55.70	25.80
quantization	71.90	47.80	24.10
400 - 800 Hz noise	83.40	61.80	21.60
1900 - 2600 Hz noise	85.10	65.00	20.10
1300 - 1900 Hz noise	86.80	68.60	18.20
BD 400 - 800	79.70	61.70	18.00
800 - 1300 Hz noise	85.80	68.90	16.90
APCM	77.70	60.90	16.80
BD 2600 - 3400	83.30	68.10	15.20
2600 - 3400 Hz noise	84.40	69.40	15.00
LPC	83.00	69.50	13.50
broadband additive noise	85.10	73.90	11.20
ECHO	87.60	76.40	11.20
0 - 400 Hz noise	86.20	75.10	11.10
lowpass filtering	85.10	74.10	11.00
BD 100 - 400	91.60	80.80	10.80
VEV 7	76.90	66.20	10.70
ADPCM	78.50	67.90	10.60
PD 1900 - 2600, radial	87.20	76.80	10.40
BD 100 - 3500	73.40	63.50	9.90

Table 5.4.1-1 Distortions which most prominently excite subjective quality SD, listed in order of decreasing significance.

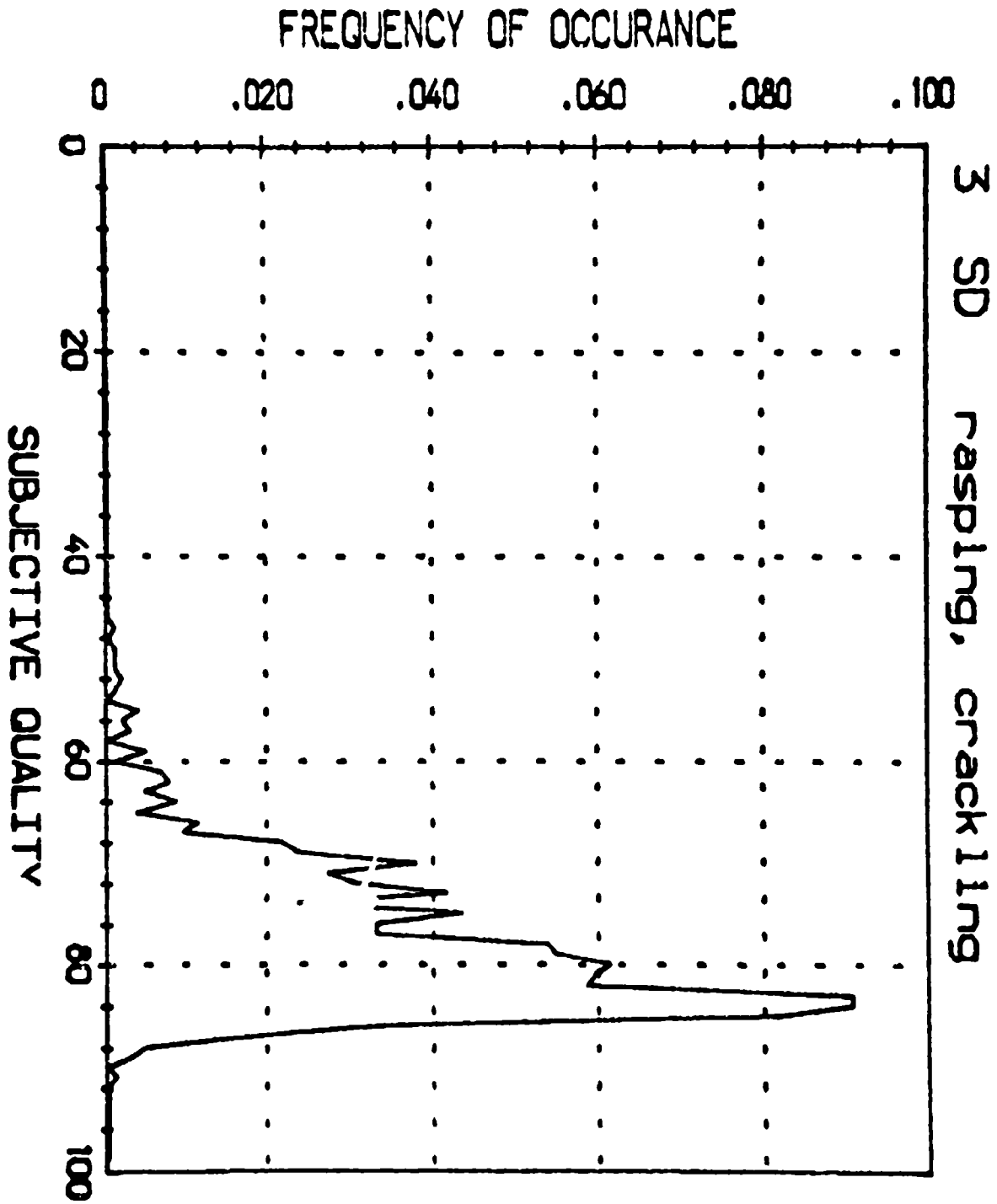


Figure 5.4.1-1 Histogram showing the value of subjective quality SD (x-axis) vs. the frequency of occurance of the SD subjective quality value (y-axis).

this issue. The horizontal axis is the SD subjective quality score. A subjective quality of 85 is very good, or nearly complete absence of the quality SD, while a 20 is very poor, or highly distorted. The vertical axis is the frequency of occurrence of a given value of the SD quality score when taken over all speakers and all distortion systems in the data base. A case by case examination of the data in this histogram would show that points which fall in the left tail of the distribution are members of the distortions listed in Table 5.4.1-1.

Research efforts up to this point have been unable to identify a good measure for this subjective speech quality. Efforts to measure the energy of the noise resulting from the nonlinear speech distortions have been largely unsuccessful because the noise energy is dominated by the speech energy. Because of this, calculating the noise power in a straightforward manner, such as by taking the difference between the power spectrums of the distorted and the original speech, is extremely prone to error.

Experiments thus far, however, indicate that a good measure for estimating SD might be some function of the difference between the level of the noise floor and the level of the excitation spectrum in a voiced segment of the distorted speech spectrum. The spectrum of an undistorted voiced speech frame is characterized by an impulsive spectrum due to the voiced excitation with a slowly varying envelope due to vocal tract filtering. The quantity to be measured, which could be called correlated SNR, is the difference between the level of a pitch peak and its adjacent valley, where both levels are measured on a log scale. The motivation for measuring this quantity is that speech which is distorted by a nonlinearity will have a slightly smeared spectrum and hence will have the difference between these two levels diminished. An objective measure for

estimating SD could be based on the correlated SNR of the distorted speech, summed over all speech frames, normalized by the correlated SNR of the original speech, also summed over all speech frames.

5.4.2 SL: Muffled, Smothered

This subjective quality describes the extent to which speech is muffled or smothered. Table 5.4.2-1 lists the distortions which excite this subjective scale. Most prominent of these is the low pass distortion which, since it eliminates high frequencies, fits well with the subjective quality of muffled. The low band bandpass distortions also produce a similar muffled quality. The other distortions fit better with the subjective quality of smothered. The highpass and the high bands of the bandpass bandlimiting distortions eliminate or diminish speech energy in the middle of the zero to 3600 Hz speech band which, produces the perceptual effect of smothered. The two waveform coders, CVSD and ADM also diminish the mid-band energy of the coded speech with respect to the original speech and hence produce the same smothered effect. The remaining two distortions listed in Table 5.4.2-1 are narrow band additive noise, both injecting noise in the low to middle part of the speech spectrum. These distortions can be thought of as smothered in that they produces a noise masking of the speech.

Like the subjective quality SD, SL has a relatively diverse mix of distortions which excite it. There are, however, far fewer types of distortions which produce severe SL quality degradations. This can be seen from the relatively small number of entries in Table 5.4.2-1 and from Figure 5.4.2-1. This Figure shows the frequency of occurrence of a specific level of the quality SL across the ensemble of all distortions. It is strikingly different from the corresponding Figure for SD in that the main lobe for quality SL is narrower and its left tail is longer and lower. This indicates

SL muffled, smothered

DISTORTION	MAX	MIN	RANGE
lowpass filtering	83.20	46.30	36.90
CVSD	87.50	62.40	25.10
bandpass filtering	77.60	53.40	24.20
ADM	87.10	68.30	18.80
center clipping	84.10	66.70	17.40
highpass filtering	79.20	62.40	16.80
400 - 800 Hz noise	85.50	69.20	16.30
800 - 1300 Hz noise	86.20	73.00	13.20

Table 5.4.2-1 Distortions which most prominently excite subjective quality SL, listed in order of decreasing significance.

Multiple R	.7342	Standard error of estimate	3.5679
Multiple R square	.5391		

Analysis of Variance

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Regression	15142.	14	1081.	84.
Residual	12946.	1017	12.	

Table 5.4.2-2 Summary of regression model used to estimate subjective quality SL.

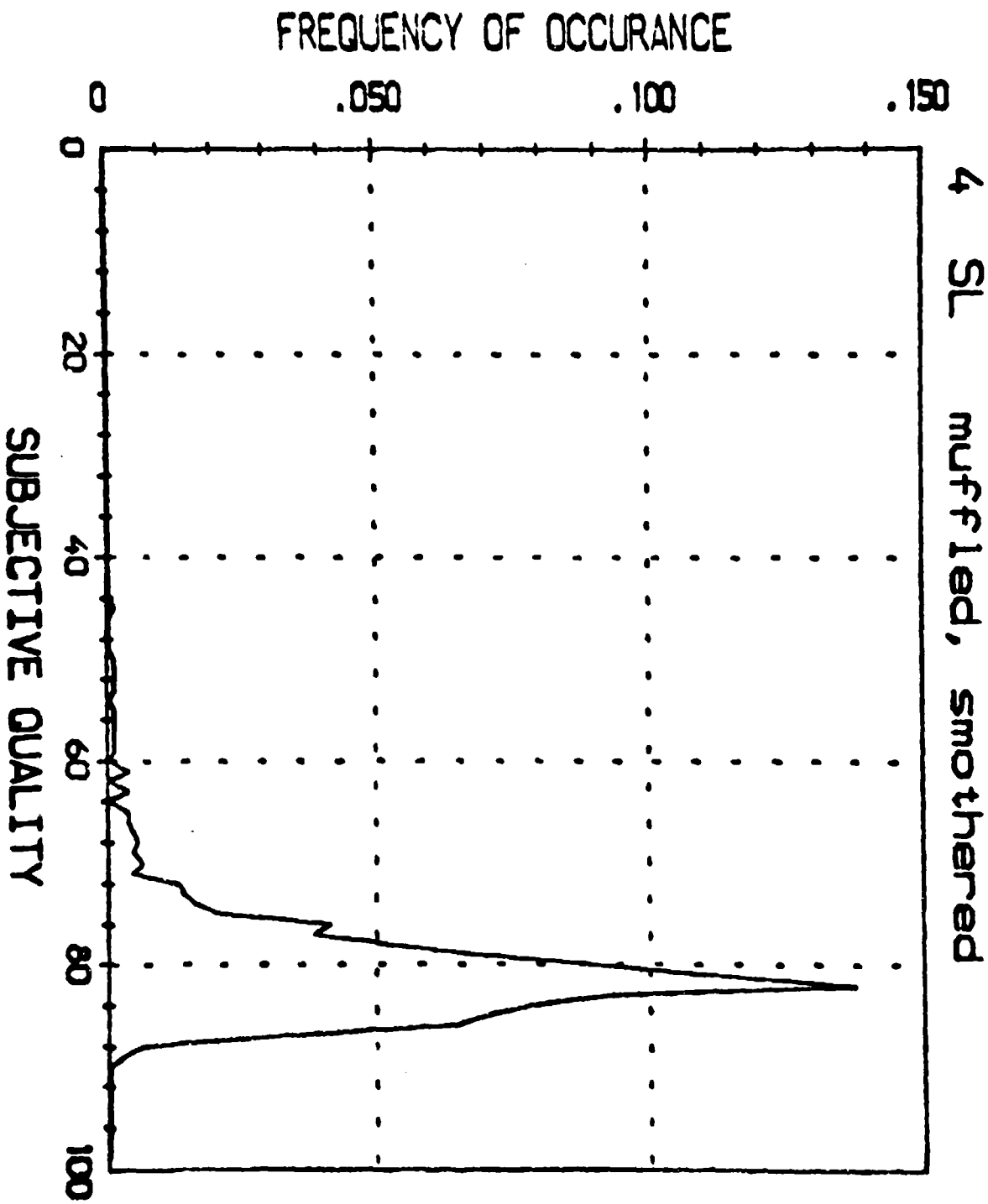


Figure 5.4.2-1 Histogram showing the value of subjective quality SL (x-axis) vs. the frequency o occurance of the SL subjective quality value (y-axis).

that the same range of quality degradation is provided by fewer distortion types.

There are primarily two types of distortions which excite the subjective scale SL. These are bandlimiting distortions and narrowband noise distortions. This suggests that a composite objective measure would be most appropriate for tracking subjective quality SL. The objective measure tried has, for its first component, a frequency variant spectral distance measure and, for its second component, a frequency variant noise measure. An important point as yet unmentioned concerning SL is that the bandlimiting and additive noise distortions which exemplify SL are time invariant systems. Therefore their distortion characteristics should be recoverable from the time averaged spectrum of the reference and distorted speech waveforms. The method used to estimate the spectrum of the waveforms was to pass the waveform through a filter bank and compute the mean square value of each filter output for each utterance. This is the same critical band filter bank used for studying aural based objective measure in Chapter 4. In this way an estimate of the power in frequency bands for an entire utterance is obtained. The power in bands could be combined, as appropriate, to provide coarser estimates of the reference and distortion spectrum. Broader bands were found to produce more easily interpreted objective measures.

The spectral distance objective measure has the following form:

$$O1(s,d,k) = \log_{10}\left(\mathrm{MIN}\left(\mathrm{MAX}\left(\frac{V(:,s,d,k)}{V(:,s,\phi,k)}, TH_{min}\right), TH_{max}\right)\right) \qquad 5.4.2\text{-}1$$

In the preceding equations, $V(.,s,d,k)$ and $V(.,s,\phi,k)$ are the mean square values in the band k for the distorted and reference waveforms, respectively. Again, this average is taken over the entire utterance. Th_{min} and TH_{max} are parameters of the measure. The objective variables $O1(s,d,k)$ were then

transformed into a new distance variable, O1*(s,b,k), which has coarser frequency resolution. Instead of having 25 bands O1* had only five bands, and is obtained by summing O1(s,d,k) as follows:

O1*(s,d,k) Band No.	O1(s,d,k) Band No.
1	1 - 5
2	6 - 10
3	11 - 15
4	16 - 20
5	21 - 25

In addition, a monotonic and uni-modal regression was done on the function O1(s,d,k) and stress for the functional forms lowpass, highpass, bandpass and band reject was computed. Computing stress for the functional form of lowpass requires computing a monotonically increasing regression, highpass requires a decreasing regression. Bandpass requires computing a global maximum uni-modal regression and band reject requires a global minimum uni-modal regression. The motivation for computing these stresses was to measure the extent to which the distortion applied to the speech had one of these bandlimiting functional forms. The total number of independent variables used this objective measure was seven: five spectral distance variables for five frequency bands and two stress variables, one for the functional form lowpass, represented as O1*(s,d,6), and one for bandpass, represented as O1*(s,d,7). The remaining stress variables did not significantly contribute to the regression model.

The second part of the composite measure is an additive noise measure. The functional form of this measure is as follows:

$$O2(s,d,k) = \log_{10}\left((1/NF) \sum_{f_S} V(f,s,d,k) + 1 \right) \qquad 5.4.2\text{-}2$$

where f_S are all silent frames in the reference utterance and NF is the number of silent frames. Like the spectral distance measure, the 25 bands in O2(s,d,k) are combined to form five bands in a new additive noise measure, O2*(s,d,k). The five variables in this measure are the noise power in the extended bands as measured during intervals of known speech inactivity in the distorted signal.

The two measures were combined in a linear function with weights determined by regression analysis. The resultant measure was formulated as:

$$O_{SL}(s,d) = \beta_0 + \sum_{j=1}^{7} \beta 1_j O1^*(s,d,j) + \sum_{j=1}^{5} \beta 2_j O2^*(s,d,j) \qquad 5.4.2\text{-}3$$

where $O_{SL}(s,d)$ represents the objective estimate of the subjective quality SL.

Table 5.4.2-2 shows the results of the multiple linear regression analysis used to formulate O_{SL}. The performance of this measure is only fair, as its correlation with SL is .74, which corresponds to an explanation of only 55% of the variability in the subjective quality SL. In all probability this poor performance is due to the difficulty of modeling the diverse mix of distortions which excite SL. This was, never the less, the best objective measure for this parametric quality.

5.4.3 SF: Fluttering, Bubbling

This subjective quality quantifies the degree to which the speech signal has a fluttering or bubbling quality. Table 5.4.3-1 lists those distortions which excite the SF subjective scale. The dominant distortion in this table is by far pole distortions. The controlled pole distortions explicitly alter the original speech pole locations, while the parametric coder distortions based on an all-pole vocal tract model distort the speech pole locations in a more

SF fluttering bubbling

DISTORTION	MAX	MIN	RANGE
interrupted, period = 1024	85.50	50.90	34.60
LPC	85.30	51.10	34.20
PD 400 - 800, frequency	85.80	52.60	33.20
interrupted, period = 300	80.90	48.70	32.20
VEV 13	84.90	60.70	24.20
VEV 7	83.30	60.30	23.00
PD 1300 - 1900, frequency	87.90	66.30	21.60
PD 400 - 800, radial	83.60	63.90	19.70
APC	86.60	67.30	19.30
BD 400 - 800	83.20	64.40	18.80
ECHO	88.60	70.80	17.80
BD 2600 - 3400	79.80	62.90	16.90
BD 100 - 3500	80.40	63.90	16.50
PD 800 - 1300, frequency	84.10	68.10	16.00
PD 000 - 400, radial	88.60	72.90	15.70
PD 1300 - 1900, radial	88.60	73.10	15.50
PD 2600 - 3400, radial	87.50	72.60	14.90
center clipping	85.60	72.10	13.50

Table 5.4.3-1 Distortions which most prominently excite subjective quality SF, listed in order of decreasing significance.

complex way through modeling errors and parameter quantization. Two prominent exceptions are the first and fourth table entries: the interrupted distortions. These are understandably perceived as fluttering because their interruptions are periodic. The presence of these interrupted distortions in Table 5.4.3-1 suggests that it is the periodic quality of the controlled and coder pole distortions which correlate most highly with subjective fluttering and bubbling.

Though it is clear that the source of degradations in the subjective quality fluttering or bubbling is primarily due to LPC pole position errors, this research was unable to identify a good measure for such errors. The interrupted component of SF could clearly be estimated by the elements of the SI objective measure, but this still leaves pole position errors or, more precisely, formant frequency and bandwidth errors, to be estimated. Further experimentation needs to be done to determine the degree to which formant frequency and formant bandwidth are correlated to SF.

In order to perform such experiments one needs a means of determining formant frequency and bandwidth for a given speech frame. The vocal tract system function as derived from LPC analysis is a good starting point for finding these parameters. The difficulty in processing this smoothed spectrum is that formant frequencies correspond to local maximums of the spectrum and are therefore hard to track. One must estimate and in some sense remove the global spectral tilt before attempting to estimate formant frequencies. Once the formants are known, calculating their bandwidths is relatively straightforward.

Once formant frequency and bandwidth can be reliably estimated, some function of the degree of variability of these parameters would seem to be a good physical correlate to subjective fluttering. One possibility is to match the first three formants in the original and the

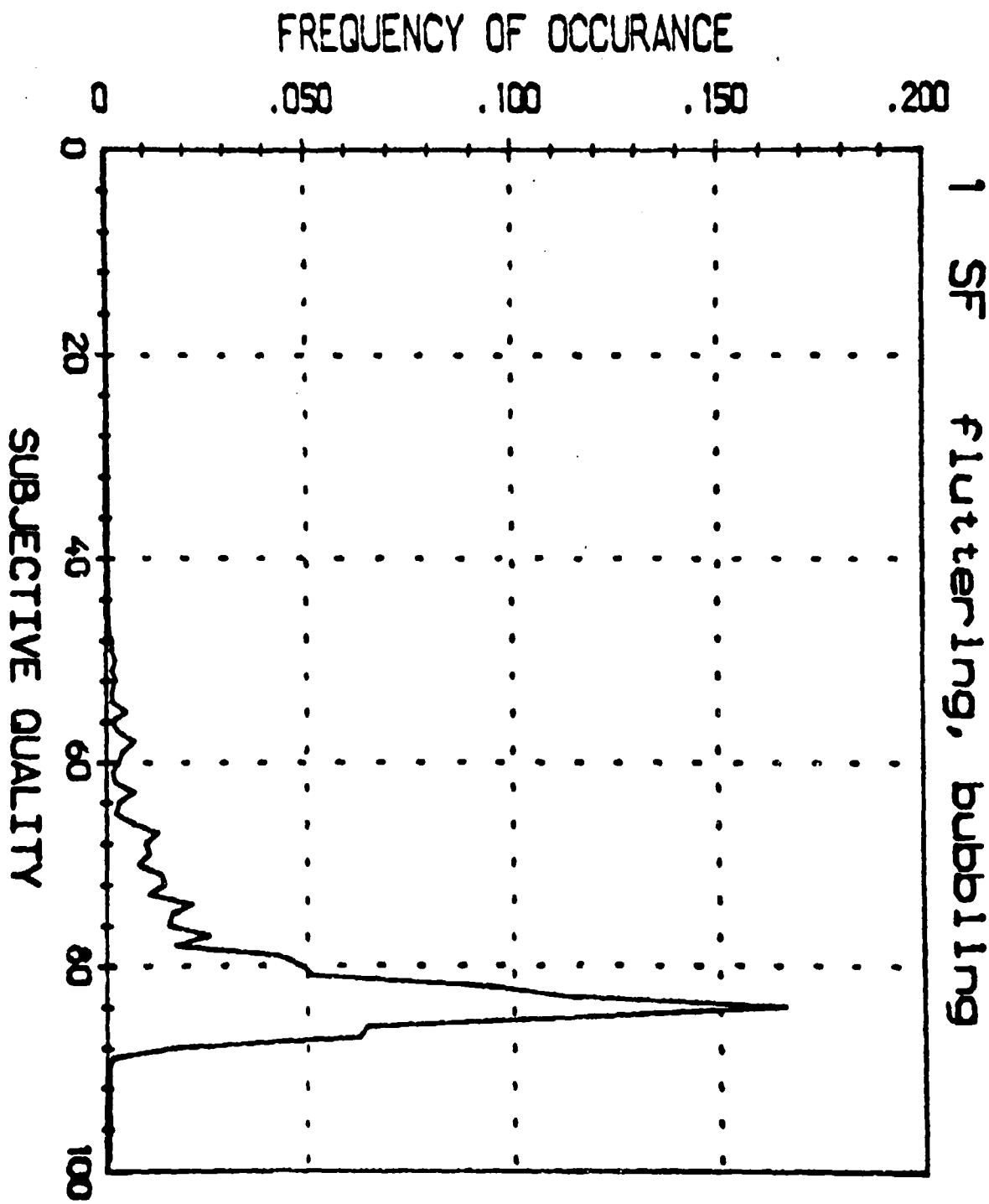


Figure 5.4.3-1 Histogram showing the value of subjective quality SF (x-axis) vs. the frequency of occurance of the SF subjective quality value (y-axis).

distorted speech frames and to calculate the variance of the difference between the distorted and original formant frequencies for each of the three pairs. The variance would be computed over the set of all speech frames. The same calculation could be done for formant bandwidth. These six objective measure variables would then be the basis for an objective measure for estimating SF.

5.4.4 BN: Hissing, Rushing

This scale specifies the extent to which the background of the distorted signal has a hissing or rushing quality. Table 5.4.4-1 lists those distortions which most excite the BN subjective scale. This scale is in contrast to the ones discussed thus far in that a very homogeneous set of controlled distortions excite this subjective quality, namely additive noise distortions. The middle frequency narrowband additive noise distortions have the greatest perceptual impact, with the broadband additive noise being perceived as almost the same degree of distortion. At the bottom of the table is quantization distortion which is not an anomaly since, for medium to fine quantization levels, the quantization noise is nearly uncorrelated with the signal and is understandably perceived as a background process.

From the evidence of the distortions which excite the BN subjective scale, a function which measures additive noise would be an appropriate objective measure for this scale. The objective measure used is that of equation 5.3.2-2, but here it is used by itself to estimate BN. The measure O2(s,d,k) is transformed into O2*(s,d,k) in order to consolidate the number of bands. The transformation is as follows:

O2*(s,d,k) Band No.	O2(s,d,k) Band No.
1	1 - 5
2	6 - 16

Note that bands 17 through 25 were not used in this measure. The objective

BN hissing, rushing

DISTORTION	MAX	MIN	RANGE
800 - 1300 Hz noise	80.40	49.30	31.10
broadband additive noise	83.40	54.00	29.40
400 - 800 Hz noise	79.10	50.40	28.70
0 - 400 Hz noise	85.80	66.40	19.40
1300 - 1900 Hz noise	82.10	69.60	12.50
2600 - 3400 Hz noise	87.20	74.80	12.40
1900 - 2600 Hz noise	84.00	72.80	11.20
quantization	85.30	75.60	9.70

Table 5.4.4-1 Distortions which most prominently excite subjective quality BN, listed in order of decreasing significance.

Multiple R	.9136	Standard error of estimate	2.3199
Multiple R square	.8346		

Analysis of Variance

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Regression	28598.	2	14299.	2656.
Residual	5667.	1053	5.	

Table 5.4.4-2 Summary of regression model used to estimate subjective quality BN.

final objective measure used to estimate BN was then:

$$O_{BN}(s,d) = \beta_0 + \sum_{j=1}^{2} \beta_j O2^*(s,d,j) \qquad 5.3.3\text{-}1$$

The performance of this measure is extremely good. The objective measure results are shown in Table 5.4.4-2. The primary reason for such good performance, correlation of .90, is that all distortions which excite BN are very similar and hence can be modeled well as a group. Another reason is that there are relatively few distortions which excite BN, as can be seen from the narrow central lobe and the low left tail of Figure 5.4.4-1. This means that the regression model need only account for the variance of these few distortions, and can approximate the quality scores of the other distortions with a constant. Of all parametric objective measures studied, this measure was by far the most successful.

5.4.5 BF: Chirping, Bubbling

This subjective quality quantifies the degree to which the speech background has a chirping or bubbling quality. Table 5.4.5-1 lists those distortions which excite the BF subjective scale. This scale is very similar to SF, or signal fluttering and bubbling. The principal differences are, first, that interrupted does not excite BF where it was at the top of the list for SF. This is understandable since an interruption of the speech waveform is a distortion entirely associated with the speech signal and produces no spurious or uncorrelated background distortion. The second difference is that high band narrowband noise distortions excite the BF scale, where they did not excite SF. These distortions are most likely perceived as chirping background distortions. The rest of the distortions listed in Table 5.4.5-1 are for the most part the same distortions associated with SF, listed in Table 5.4.3-1. Therefore an

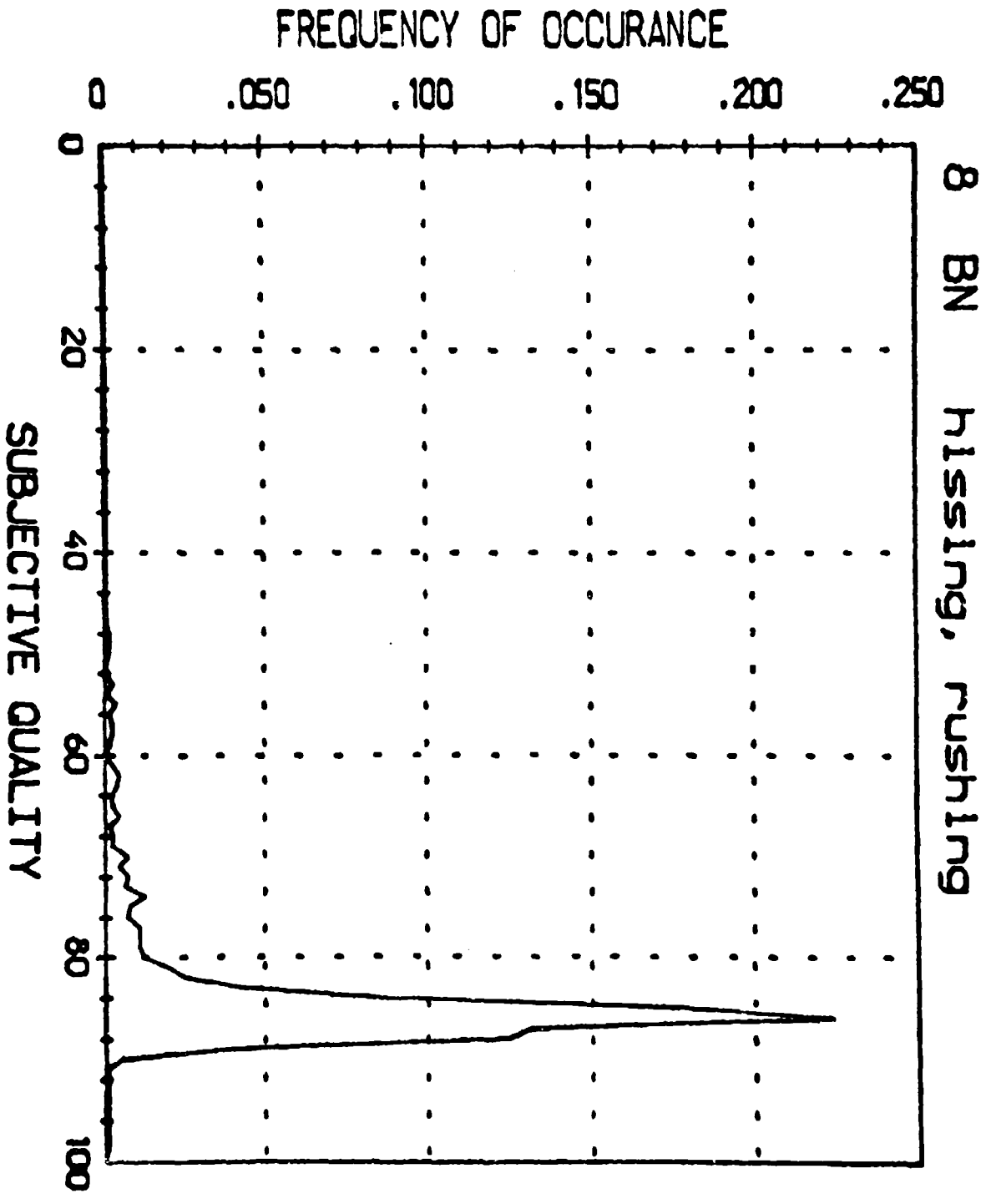


Figure 5.4.4-1 Histogram showing the value of subjective quality BN (x-axis) vs. the frequency of occurance of the BN subjective quality value (y-axis).

BF
chirping,
bubbling

DISTORTION	MAX	MIN	RANGE
PD 1300 - 1900, radial	85.70	54.40	31.30
PD 800 - 1300, frequency	86.40	57.20	29.20
LPC	85.10	56.00	29.10
PD 1900 - 2600, radial	85.10	57.90	27.20
PD 400 - 800, radial	85.30	59.20	26.10
PD 400 - 800, frequency	85.60	59.60	26.00
PD 000 - 400, radial	85.20	59.60	25.60
PD 800 - 1300, radial	87.50	65.90	21.60
PD 1300 - 1900, frequency	86.90	66.30	20.60
VEV 7	77.40	59.00	18.40
VEV 13	76.20	59.90	16.30
PD 2600 - 3400, frequency	86.70	70.70	16.00
PD 2600 - 3400, radial	90.00	74.60	15.40
APC	84.40	69.10	15.30
PD 1900 - 2600, frequency	87.50	72.30	15.20
PD 2600 - 3400, frequency	87.50	73.80	13.70
BD 2600 - 3400	83.60	70.80	12.80
2600 - 3400 Hz noise	85.10	72.70	12.40
BD 100 - 3500	81.80	69.80	12.00
1900 - 2600 Hz noise	83.60	71.60	12.00
BD 400 - 800	83.50	71.80	11.70
1300 - 1900 Hz noise	86.40	74.70	11.70
BD 1300 - 1900	83.00	71.90	11.10
quantization	85.10	74.10	11.00
BD 800 - 1300	81.60	70.80	10.80

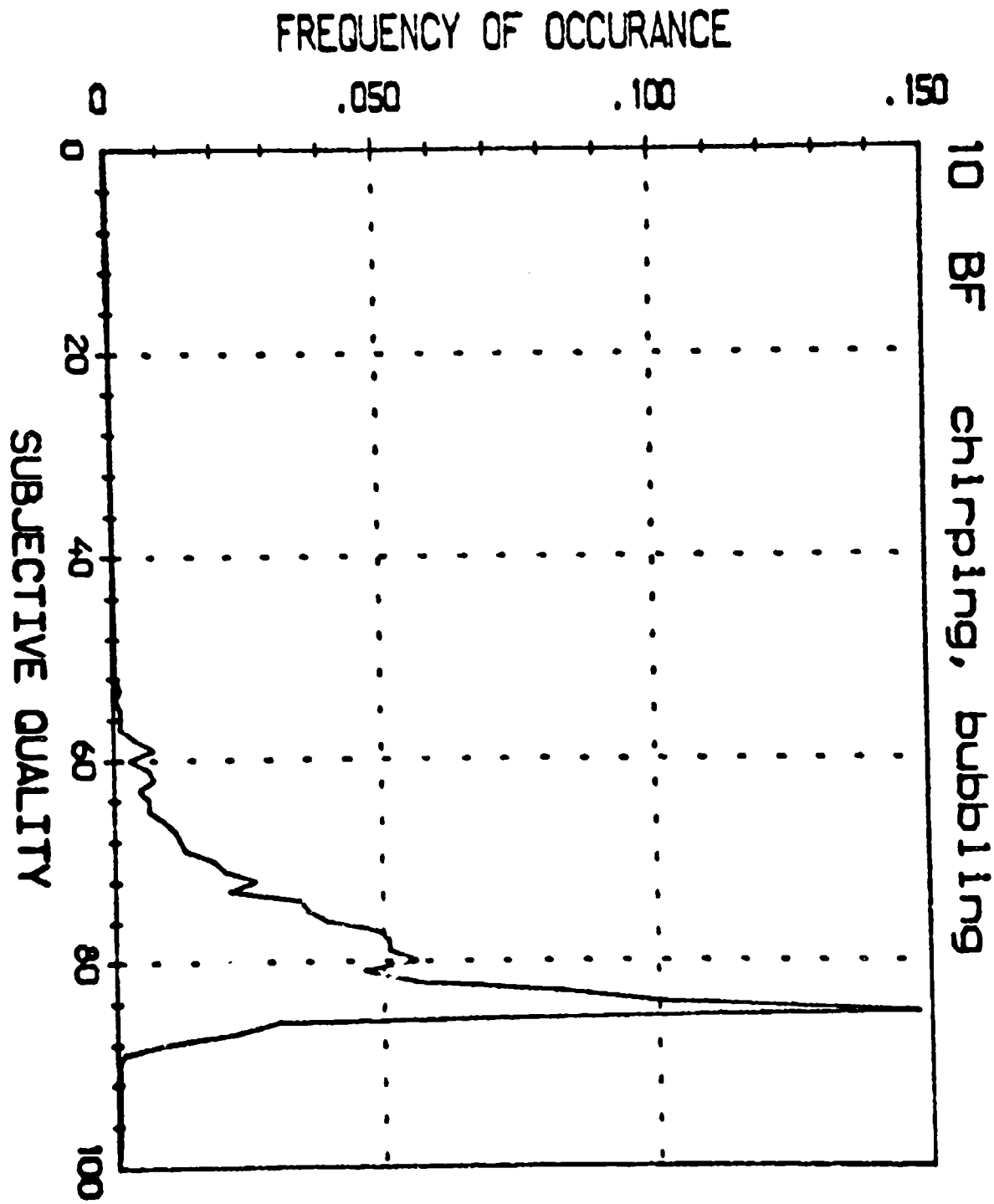


Figure 5.4.5-1 Histogram showing the value of subjective quality BF (x-axis) vs. the frequency of occurance of the BF subjective quality value (y-axis).

objective measure for estimating BF should be similar to a measure for SF. Referring back to the multidimensional scaling of the subjective data base, Figure 5.3.2-1, one can see that SF and BF are both at the bottom of the plot and are rather close together, confirming the fact that the two quality scales detect perceptually similar distortions.

This research was unable to identify good objective measures for BF. This is largely to be expected since SF was also difficult to objectively estimate. The same insight into objective measures for SF, as discussed in section 5.4.3, largely holds true for objective measures for BF. The primary difference is that objective estimates of interrupted are not needed for estimating BF while objective estimates of background noise are. The latter objective estimates are discussed in section 5.4.4.

5.4.6 SI: Irregular, Interrupted

This parametric quality scale describes the degree to which the speech signal is irregular and interrupted. Table 5.4.6-1 lists distortions which excite this subjective scale. The most prominent distortion is the slow periodic interruption, with the fast periodic interruption falling in the middle of the table. These two distortions certainly produce perceptually interrupted speech. It is difficult to find an objective quality which is common to the remainder of the distortions which excite SI. They most likely excite the subjective quality irregular, rather than interrupted. The remaining distortions are not totally disjoint, however. Both APCM and ADPCM excite SI, and the two highest bands of the narrowband additive noise excite SI. Several pole distortions also excite SI.

Since interrupted is the most important aspect of the SI scale and since this quality is easy to model objectively, the measure used for estimating SI was designed to respond only to interruptions of the speech waveform. In particular the average number of consecutive frames for which the distorted

SI irregular, interrupted

DISTORTION	MAX	MIN	RANGE
interrupted, period = 1024	87.90	38.40	49.50
ADM	91.00	49.60	41.40
2600 - 3400 Hz noise	87.10	62.50	24.60
ADPCM	85.00	60.50	24.50
center clipping	87.60	63.90	23.70
interrupted, period = 300	86.80	66.20	20.60
APCM	85.20	65.10	20.10
ECHO	89.90	76.20	13.70
PD 800 - 1300, frequency	89.50	76.50	13.00
PD 1900 - 2600, radial	89.90	77.80	12.10
PD 000 - 400, radial	89.60	79.10	10.50
PD 1900 - 2600, frequency	89.20	79.30	9.90
1900 - 2600 Hz noise	87.10	78.70	8.40

Table 5.4.6-1 Distortions which most prominently excite subjective quality SI, listed in order of decreasing significance.

Multiple R .8483 Standard error of estimate 2.6043
Multiple R square .7196

Analysis of Variance

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Regression	17454.	4	4363.	643.
Residual	6802.	1003	6.	

Table 5.4.6-2 Summary of regression model used to estimate subjective quality SI.

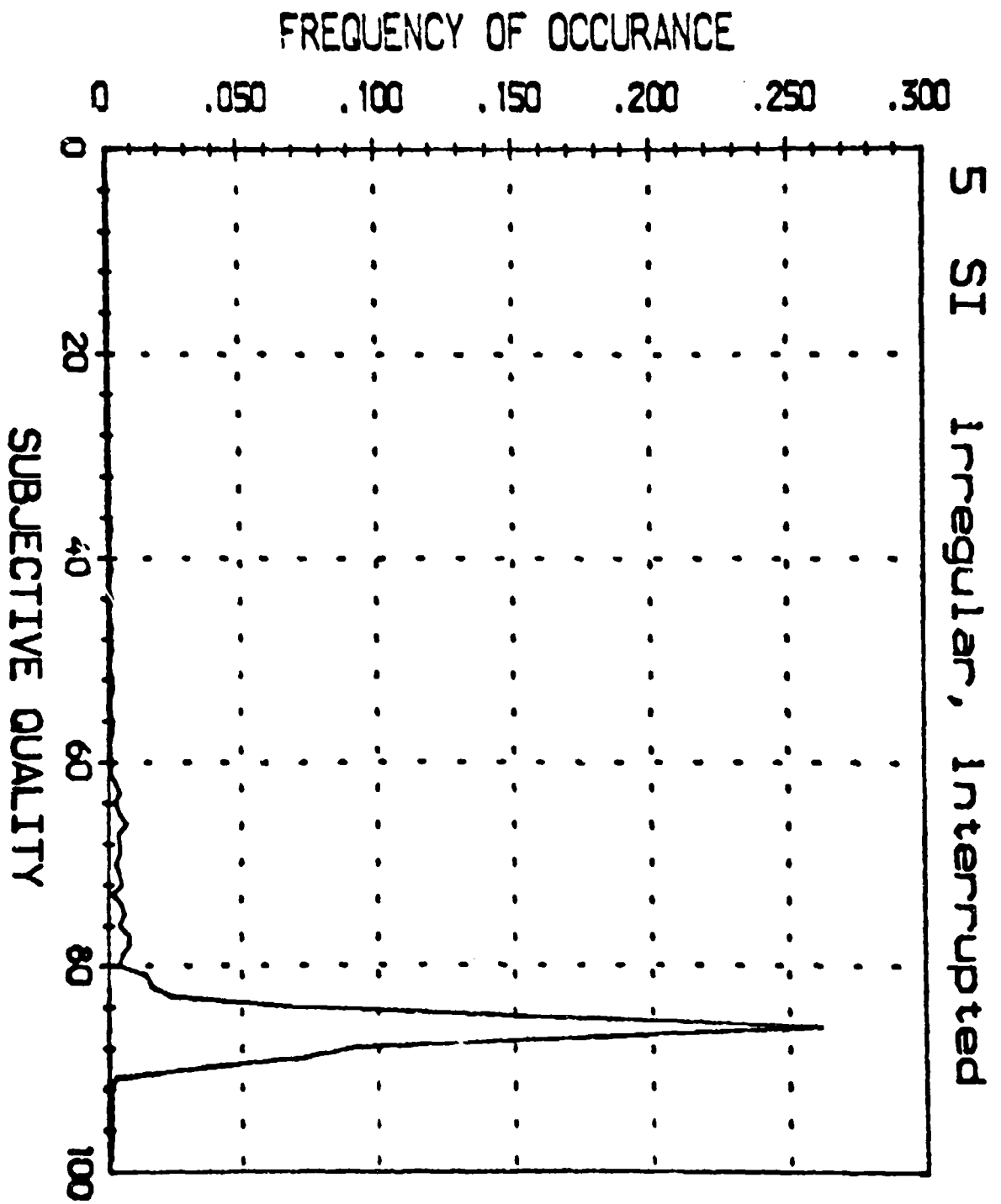


Figure 5.4.6-1 Histogram showing the value of subjective quality SI (x-axis) vs. the frequency of occurance of the SI subjective quality value (y-axis).

speech energy was either below a specified threshold or above the threshold is measured as a gauge of interruption. The measure is best expressed using intermediate variables follows:

$$RATIO(f,s,d) = \frac{\log_{10}\left(\left((1/FL) \sum_{m_f} X(m,s,d) \right)^{0.5} \right)}{\log_{10}\left(\left((1/FL) \sum_{m_f} X(m,s,0) \right)^{0.5} \right)} \qquad 5.4.6\text{-}1$$

$$ON(s,d) = \text{Average run length of frames for which } (RATIO(f,s,d) > TH) \qquad 5.4.6\text{-}2$$

$$OFF(s,d) = \text{Average run length of frames for which } (RATIO(f,s,d) < TH) \qquad 5.4.6\text{-}3$$

$$O(s,d,1) = OFF(s,d) \qquad 5.4.6\text{-}4$$

$$O(s,d,2) = \frac{OFF(s,d)}{(ON(s,d) + OFF(s,d))} \qquad 5.4.6\text{-}5$$

$$O(s,d,3) = O(s,d,1) \qquad 5.4.6\text{-}6$$

$$O(s,d,4) = O(s,d,2) \qquad 5.4.6\text{-}7$$

$$O_{SI}(s,d) = \beta_0 + \sum_{j=1}^{4} \beta_j O(s,d,j) \qquad 5.4.6\text{-}8$$

Parameters FL and TH can be varied as desired to alter the measure. Parameter FL is the number of samples in a frame of speech and parameter TH specifies the threshold between objectively interrupted and non-interrupted speech. In the formula specifying RATIO, m_f is the index of the speech samples comprising frame f. The objective measure variables are specified in equations 5.4.6-4 through 5.4.6-7. Note that the last two objective variables are simply the first two objective variables squared. Therefore the final objective measure specified in equation 5.4.6-8 is actually a multiple linear and polynomial regression equation.

The results of using regression analysis to find the best estimate, O_{SI}, of quality SI are shown in Table 5.4.6.2. The measure performed reasonably well, as measured by a multiple R of .85, with the restriction that not all the distortions were included in the analysis. Specifically, ADPCM, APCM and ECHO were not included in the analysis. ECHO was excluded because it was not representative of typical speech coder distortions. However, ADPCM and APCM were excluded because their distortions were not being modeled well by this objective measure. Leaving them out greatly improved the correlation with SI. As mentioned previously, these two coder distortions most likely produce a subjectively irregular distortion. This is, admittedly, a rather major shortcoming of this objective measure, but a future composite measure made up of this measure and another measure which does track perceived irregularity would rectify this deficiency.

5.4.7 SH: Distant, Thin

This last subjective quality measures the degree to which the distorted speech sounds distant or thin. The distortions which most dramatically excite this parametric quality scale are bandlimiting distortions, specifically highpass and bandpass distortions. These two distortions are ordered one and two in Table 5.4.7-1. For the higher bands, the bandpass filtering is very similar to highpass filtering so it is reasonable that these two distortions are grouped together. They indicate that highpass filtering is the most important objective correlate to speech being perceived as distant and thin. Two seemingly out of place distortions found in Table 5.4.7-1 are CVSD and lowpass filtering. On closer inspection CVSD does in fact produce a bandlimiting distortion which slightly decreases the energy of speech in a broad band centered at approximately 2000Hz. So the only feature these two distortions have in common is that they both diminish speech energy in mid band, although lowpass filtering eliminates virtually all out of band energy. A

SH distant, thin

DISTORTION	MAX	MIN	RANGE
highpass filtering	84.70	54.20	30.50
bandpass filtering	85.00	60.60	24.40
0 - 400 Hz noise	86.90	75.40	11.50
CVSD	90.30	79.30	11.00
lowpass filtering	87.90	78.00	9.90
peak clipping	87.10	79.60	7.50

Table 5.4.7-1 Distortions which most prominently excite subjective quality SH, listed in order of decreasing significance.

Multiple R .8540 Standard error of estimate 2.4545
Multiple R square .7293

Analysis of Variance

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Regression	17023.	6	2837.	470.
Residual	6319.	1049	6.	

Table 5.4.7-2 Summary of regression model used to estimate subjective quality SH.

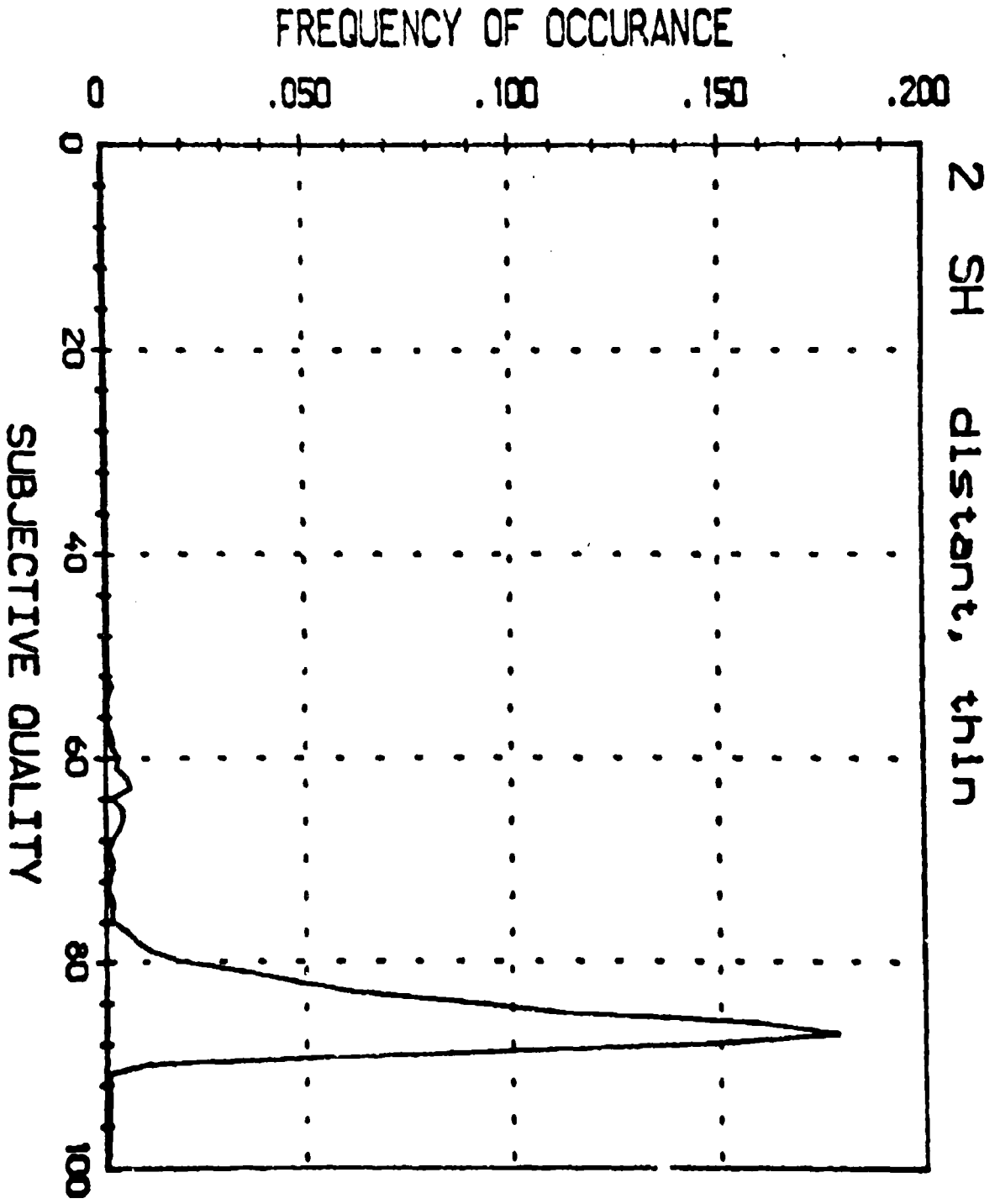


Figure 5.4.7-1 Histogram showing the value of subjective quality SH (x-axis) vs. the frequency of occurance of the SH subjective quality value (y-axis).

possible conclusion is that like SL, SH is correlated to a decrease in mid-band speech energy. Other distortions which excite SH are peak clipping and the lowest band of narrowband additive noise. Peak clipping smears energy across the entire spectrum which is perceived primarily as high frequency distortion due to the low level of speech energy at high frequencies. Therefore these two distortions produce noise at opposite ends of the spectrum. This effect may be correlated to the quality distant and thin.

The objective measure used to estimate SH concentrated on the principle objective feature of SH which is highpass filtering. The objective measure is a spectral distance measure which is identical to the one used to estimate SL, specified in equations 5.3.2-1 and including the subsequent transformation to reduce the number of bands to five. The objective distance variables are combined in a regression equation for estimating SH as follows:

$$O_{SH}(s,d) = \beta_0 + \sum_{j=1}^{5} \beta_j O^*(s,d,j) \qquad 5.4.7\text{-}1$$

Table 5.4.7-2 shows the results of this analysis. Performance for this measure was significantly better than for the measure which estimates SL. For this measure a correlation of .85 was obtained. This is primarily due to the fact that the distortions which produce most of the variance in SH, highpass and bandpass filtering, are relatively homogeneous and therefore can be effectively modeled.

5.5 Discussion

In the previous section we have presented four parametric objective measures. The performance of these measures range from very good (a correlation of 0.90 for BN) to fair (a correlation of 0.74 for SL). Though these results are quite good, they are more remarkable because the

objective measures estimated subjective quality over the entire distorted data base. (with the exception of O_{SI}.) This is encouraging because it indicates that these objective measures are applicable to a broad range of speech distortions.

Objective measures with similar performance could not be found for subjective qualities SD, SF and BF, though the probable form of measures for estimating these subjective qualities was discussed. Further analysis is necessary to better understand the physical manifestations of these perceptual qualities before good measures for them can be designed.

In designing each parametric objective measure, we have attempted to build regression models in which all of the regression weights have an intuitively satisfying physical interpretation. The ability to assign a meaning to the regression coefficients is a check on the appropriateness of the regression model. More complex models with relatively meaningless regression weights have been avoided. Even though such models are able to provide improved performance, it is suspected that they do so by accounting for variations in the noise of the data and do not provide improved modeling of the subjective speech perception process.

In some cases the parametric objective measure results may have utility by themselves. For example, a low score on the BN objective measure may indicate excessive additive noise distortion in the speech system, while a low score on the SF objective measure may indicate insufficient quantization levels in the vocal tract parameters of an LPC based speech coder. In general, the parametric measures yield specific information which may be extremely useful in diagnosing the cause of voice quality degradation in a communications system.

However, for verification of overall performance of a speech communication network, an objective measure for composite acceptability

is needed. Such a measure can be used in the design of speech communication systems and in the field maintenance of speech systems. Given that we have a full set of parametric objective measures which provide good estimates of SD, SL, SF, BN, BF, SI and SH, the essential information in these parametric measures, the objective measure variables, can be used to build a measure of composite acceptability. The form of the objective measure would be as follows:

$$O_{CA} = \beta_0 + \sum_{j=1}^{m} \beta_j O_{i,j} \qquad 5.5\text{-}1$$

where i is an index over speakers and distortion systems and j is an index over the included objective measure variables. The variables $O_{i,j}$ are the same objective variables used in constructing the parametric measures, though they are now lumped together in a single regression model and each is weighted by a β_j unique to this new model. A problem with equation 5.5-1 is that it models CA as a linear combination of the objective measure variables. This inadequacy can be lessened if interaction terms, or product terms involving the objective measure variables, are added to the model.

The key to designing a good measure for composite acceptability is to represent all significant perceptual dimensions of acceptability in the model. This point was illustrated by the multidimensional scaling analysis of the subjective data base in section 5.3.2. Because the objective measure variables used in equation 5.5-1 contain all the information needed to estimate the most significant parametric subjective qualities, they in some sense span the perceptual space of subjective composite acceptability. It is therefore reasonable to expect

that this measure for CA will perform as well as any of the individual measures of parametric subjective quality.

REFERENCES

[1] D.C. Montgomery, E.A. Peck, Introduction to Linear Regression Analysis, New York: John Wiley, 1982 Chapter 2 and 3.

[2] R.E. Barlow, Statistical Inference Under Order Restrictions, New York: John Wiley, 1972, pp. 1-27.

[3] R.N. Shepard, 'The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function,' Psychometria, vol. 27, pp 125-139, pp. 2129-246, 1962.

[4] R.N. Shepard, Multidimensional Scaling: Theory and Applications in the Behavioral Sciences, vol. 1, New York: Academic Press, 1972, pp. 1-44, 105-153.

[5] J.B. Kruskal, 'Multidimensional Scaling by Optimizing Goodness of Fit to a Numerical Hypothesis,' Psychometria, vol. 29, pp. 1-27, 1964.

[6] J.B. Kruskal, 'Nonmetric Multidimensional Scaling: a Numerical Method,' Psychometria, vol. 29, pp. 115-129, 1964.

[7] S.S. Schiffman, Introduction to Multidimensional Scaling, New York: Academic Press, 1981.

[8] J.B. Kruskal, Multidimensional Scaling, Beverly Hills: Sage Press, 1978.

CHAPTER 6

PRECLASSIFIED OBJECTIVE SPEECH QUALITY MEASURES

6.1 Introduction

In the previous two chapters, two distinct approaches to the design of objective speech quality measures were studied in some detail. Chapter 4 studied the use of aural models in designing objective measures while Chapter 5 studied the use of parametric objective measures for the same purpose. Both of these approaches met with some degree of success. This chapter introduces and develops yet another separate approach to designing objective quality measures: that of preclassifying (or labeling) the distortions before the application of the objective measures. The basic procedure used in this approach has three steps. In the first of these, each speech distortion to be measured is assigned to a specific class of distortions. This classification may be done either objectively or subjectively, although objective classification is much more desirable. Once all of the distortions are classified, then separate objective measures are designed for each separate class of distortion. Finally, these separate classified objective measured are combined to form a single, broadly based objective measure.

It is simple to show that the preclassification of distortions leads to vast variations in the performance of simple objective measures. Figure 6.1-1 shows a plot of the correlation coefficient for a log spectral distance measure as a function of the value of p in the L_p norm [6.1]. The results are shown separately for the cases in which the objective measure is applied to all distortions in the distorted data base, and three distortion subsets: controlled distortions, waveform coders, and all coders. Clearly, the log spectral distance measure performs much better on some of these distortions than on others. The point here is that if the distortions could be correctly

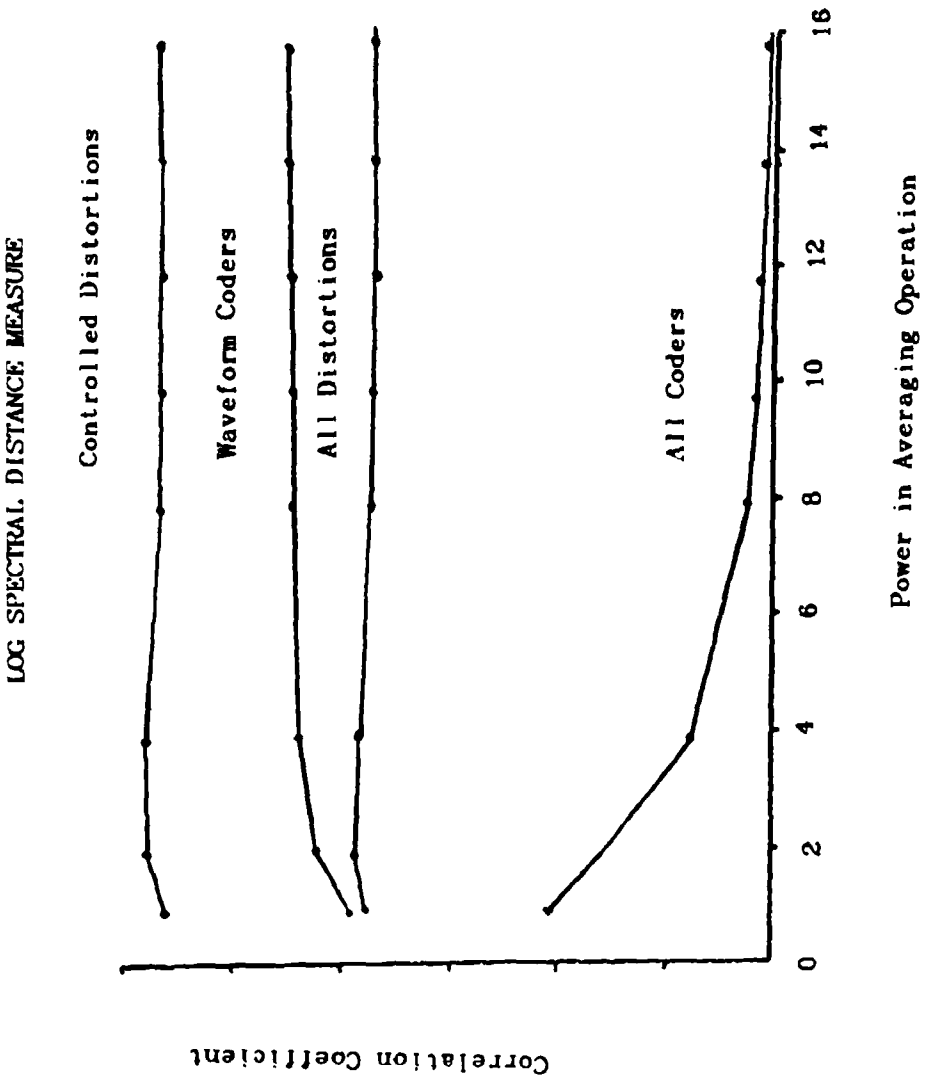


Figure 6.1-1 Plot of Log Spectral Distance Measures as a Function of p in the L_p Norm for Four Different Distortion Classes

classified, then objective measures which had been specifically designed for the proper class could be applied, resulting in better overall performance.

6.2 Objective Measures for Narrow Distortion Classes

There are really two questions to be addressed here. The first question is given a good measure for classifying measures, what is the expected improvement in the overall performance of the objective measures. If the performance improvement is small, then there is no need for more extensive study. If the answer to the first question is positive, then the second question is how to assign objectively a particular distortion to a particular class in order to realize the expected improvement.

Figures 6.2-1 and 6.2-2 show the composite acceptability (CA) results for the the six distortion levels of CVSD and APC respectively. In both cases, the results are displayed parametrically as a function of talker. There are two points which should be noted from these figures. First, for each individual talker, these results could be well represented by a first or second order regression model. Second, there is a considerable and consistent spread of results between the talkers. Hence, subjective measure results from one talker are not necessarily good predictors of subjective measure results from another talker. Clearly, a good classified objective measure must also exhibit this talker selectivity if it is going to be a good predictor of subjective responses.

Figures 6.2-3 and 6.2-4 illustrate the use of narrowly classified simple objective measures for CVSD and APC. The measures illustrated on these plots include the log spectral distance measure with linear regression, the log spectral distance measure with non-linear regression, and the short-time frequency variant SNR. Clearly, the performance of these simple measures is substantially improved by the classification process.

Figures 6.2-5 and 6.2-6 illustrate the use of narrowly classified

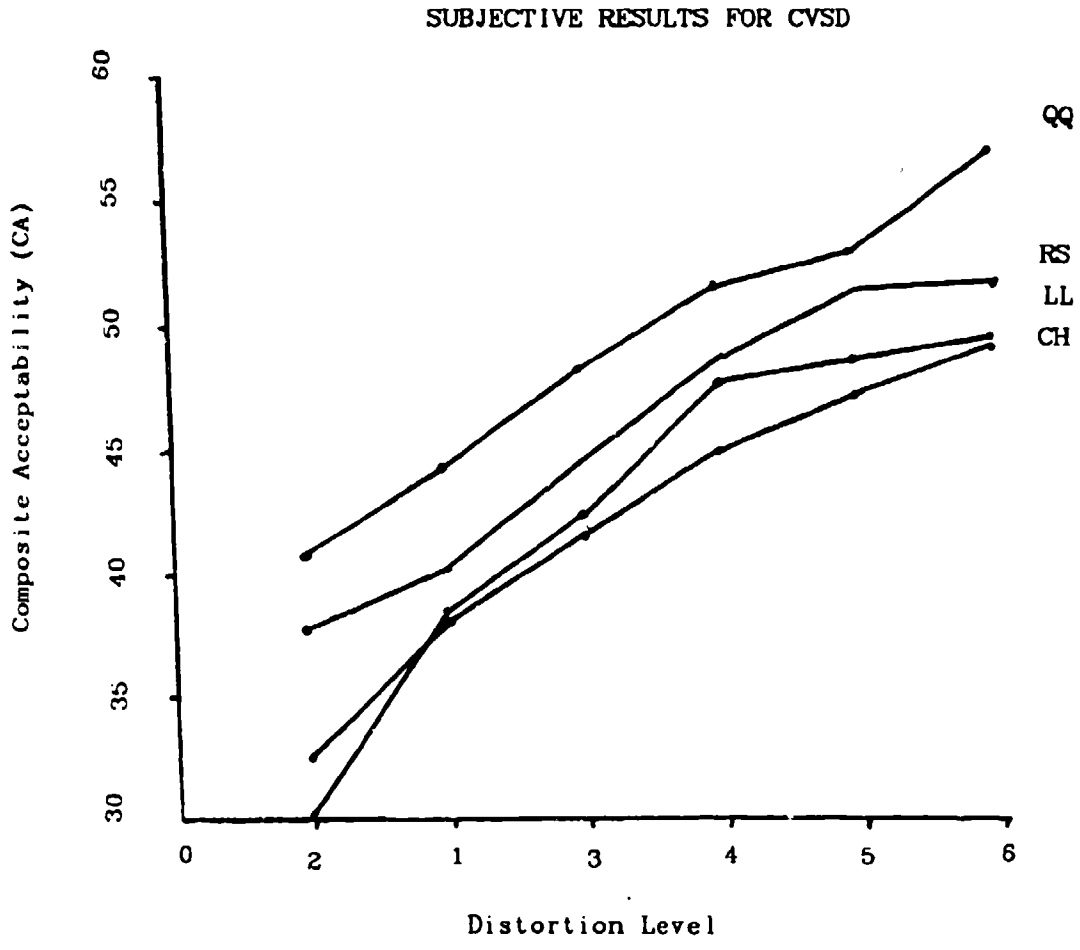


Figure 6.2-1 Composite Acceptability for CVSD as a Function of Talker and Distortion Level

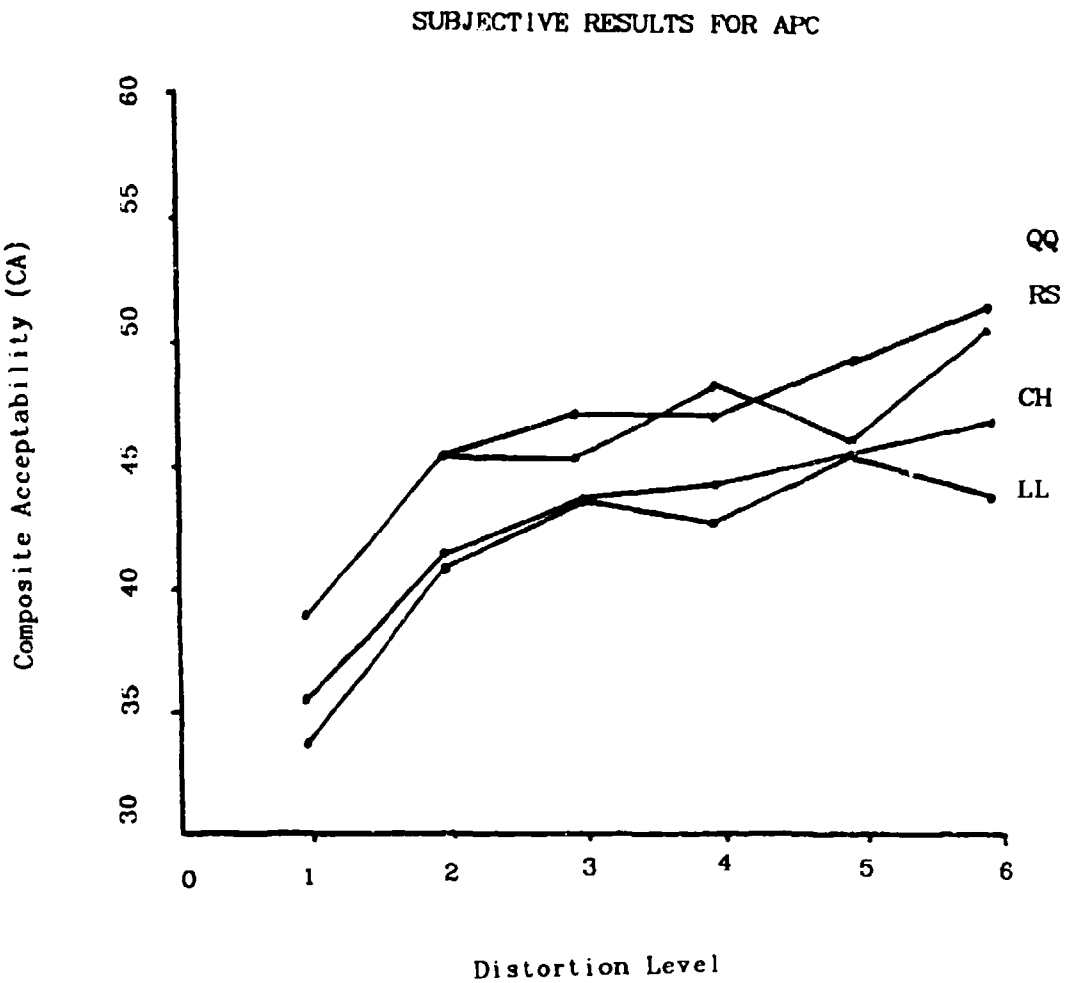


Figure 6.2-2 Composite Acceptability for APC as a Function of Talker and Distortion Level

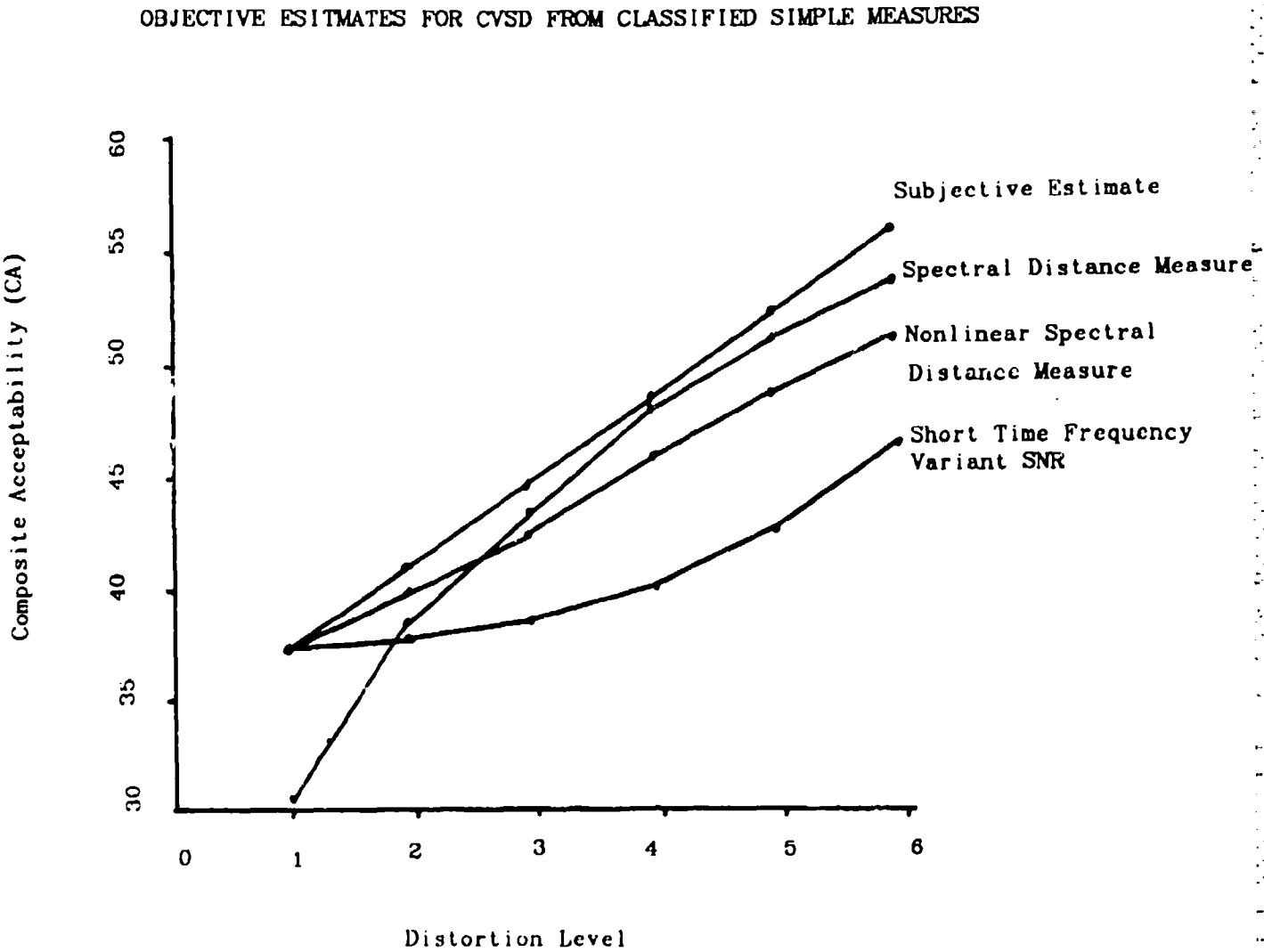


Figure 6.2-3 Objective Estimates for Composite Acceptability (CA) for CVSD from Simple Classified Objective Measures

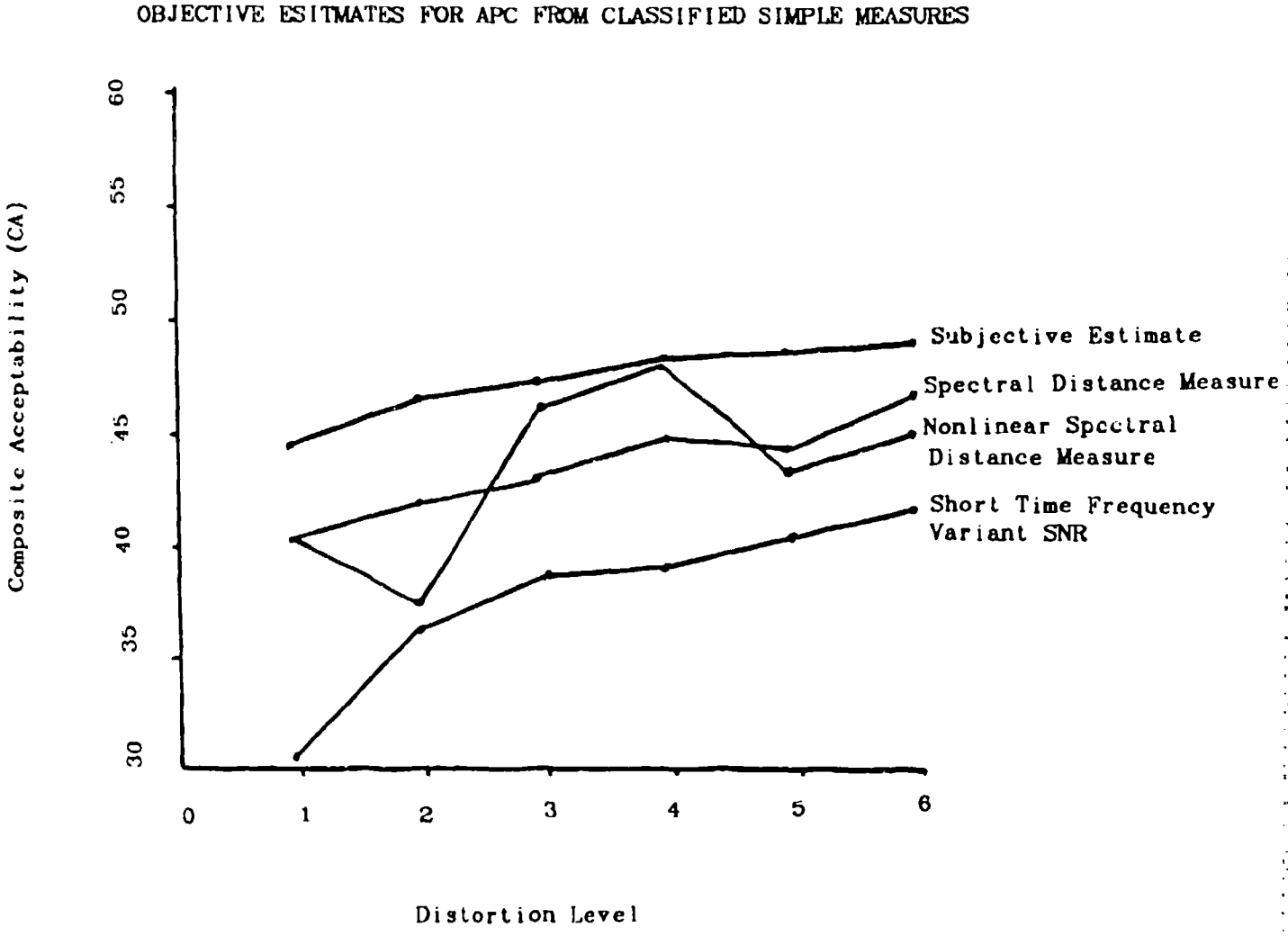


Figure 6.2-4 Objective Estimates for Composite Acceptability (CA) for APC from Simple Classified Objective Measures

composite objective measures for CVSD and APC. In each case, the measure used was trained specifically to predict only the distortions in the two classes. The objective measures used were the short-time frequency-variant SNR, a linear multi-regression composite measure, and a non-linear multi-regression composite measure [6.1]. As can be seen from these plots, the performance of each of the narrowly defined objective measures is, on the whole, very good. Indeed, a comparisons with Figures 6.2-1 and 6.2-2 show that these narrow objective measures are better predictors of CA than individual one-talker subjective measures. Figures 6.2-7 and 6.4-8 illustrate the reason for this good performance. These figures show the objective and subjective estimates of composite acceptability for the linear composite measure as a function of individual talker. Clearly, this measure has good talker selectivity.

Based on the above discussion, it is possible to make two general statements. First, if the class of distortions of interest are narrow enough, then it is possible to design composite measures which predict the subjective quality with remarkable accuracy. This is an important fact if the goal is to determine if a known coding system is performing up to standard and to measure the level of the reduced performance if it is not. Second, if the class of distortions of interest is broad, then the required task is to classify the candidate into a narrow class so as to gain the advantage discussed above. So the fundamental question reduces to finding procedures to classify distortions objectively.

6.3 Identification of Homogeneous Subsets in the Distorted Data Base

6.3.1 Introduction

There are two broad approaches to searching for improved objective speech quality measures. The first is to find measures which provide improved quality estimates over a broad range of distortions. The second is to find measures which provide improved quality estimates over a restricted range of

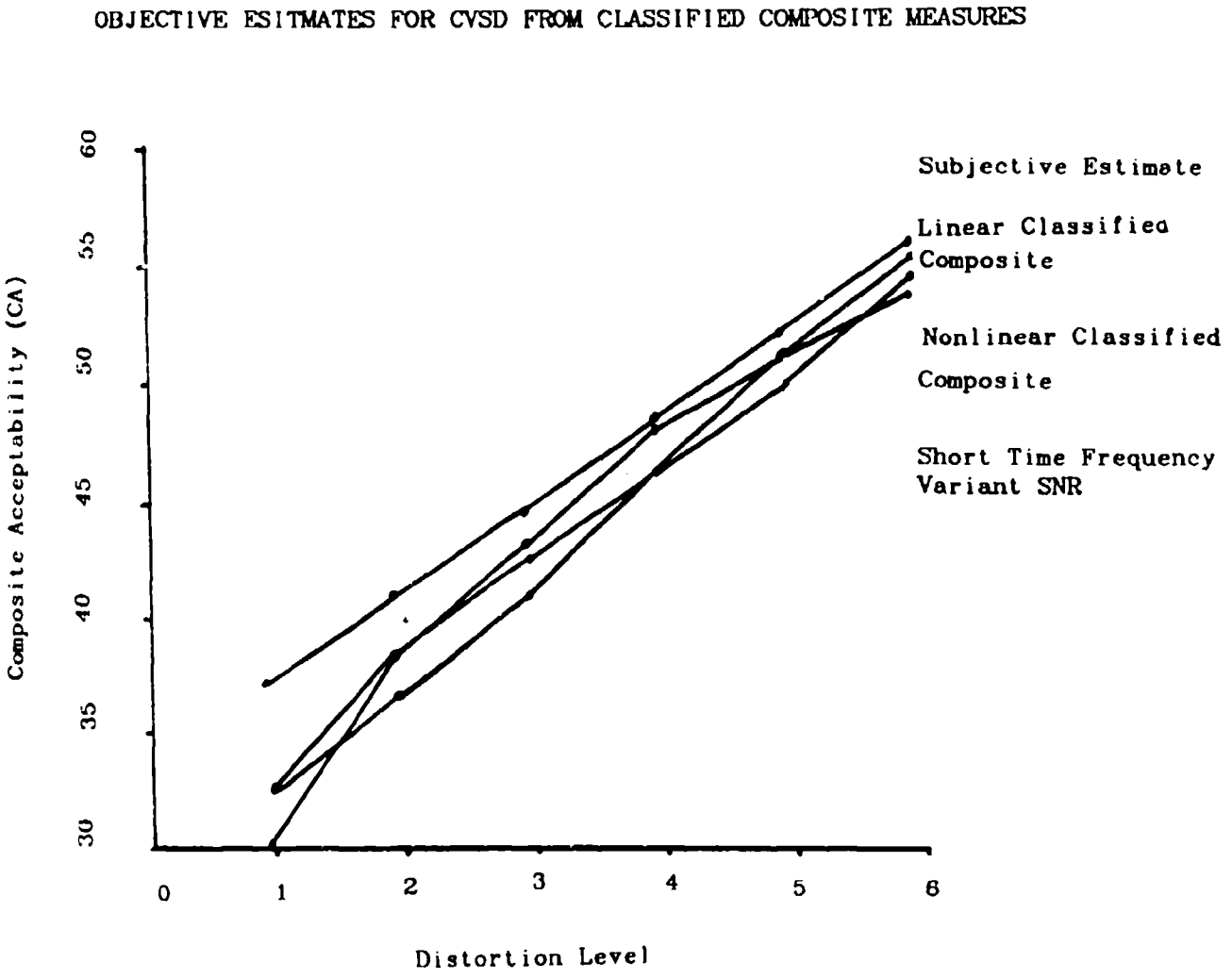


Figure 6.2-5 Objective Estimates for Composite Acceptability (CA) for CVSD from Composite Classified Objective Measures

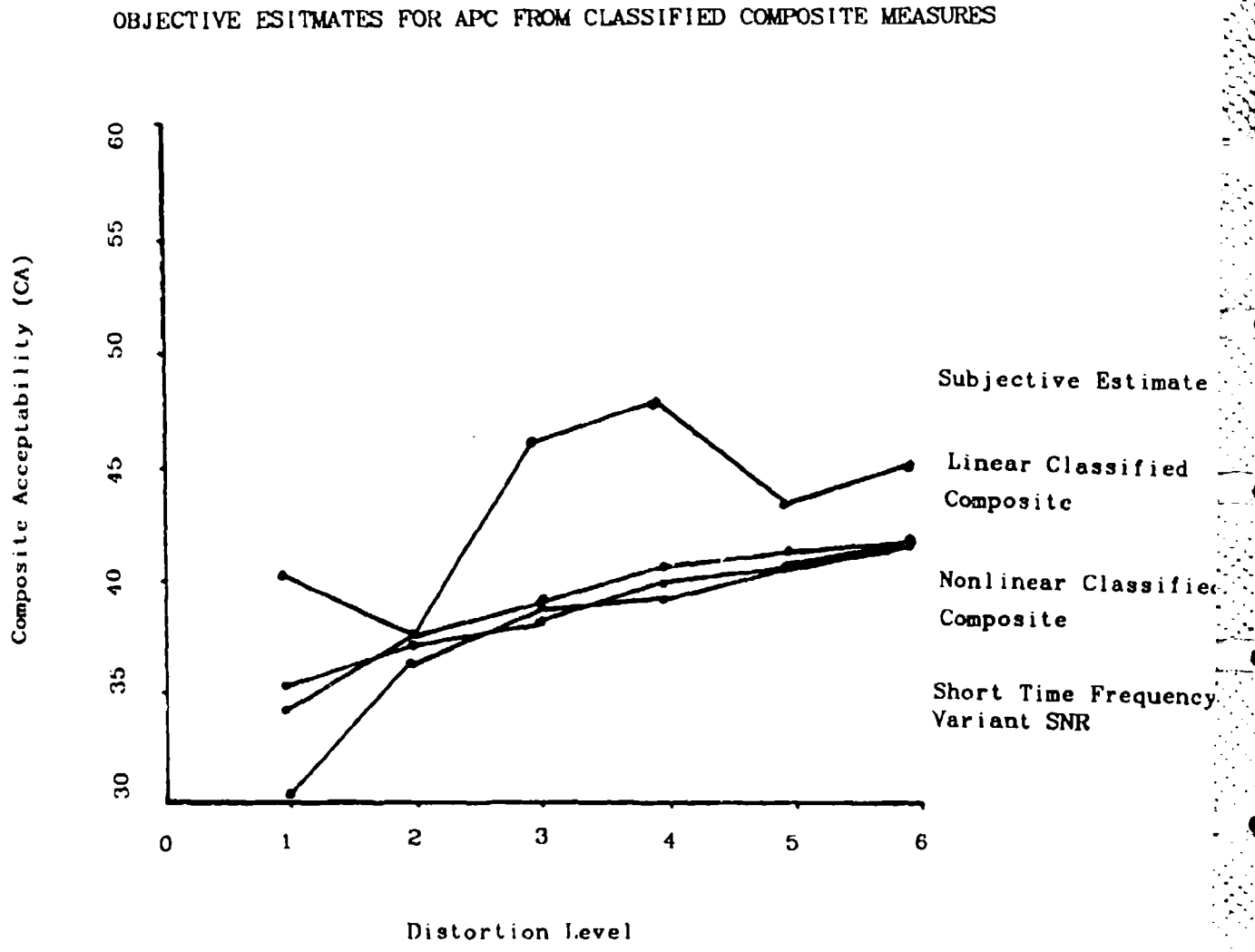


Figure 6.2-6 Objective Estimates for Composite Acceptability (CA) for APC from Simple Classified Objective Measures

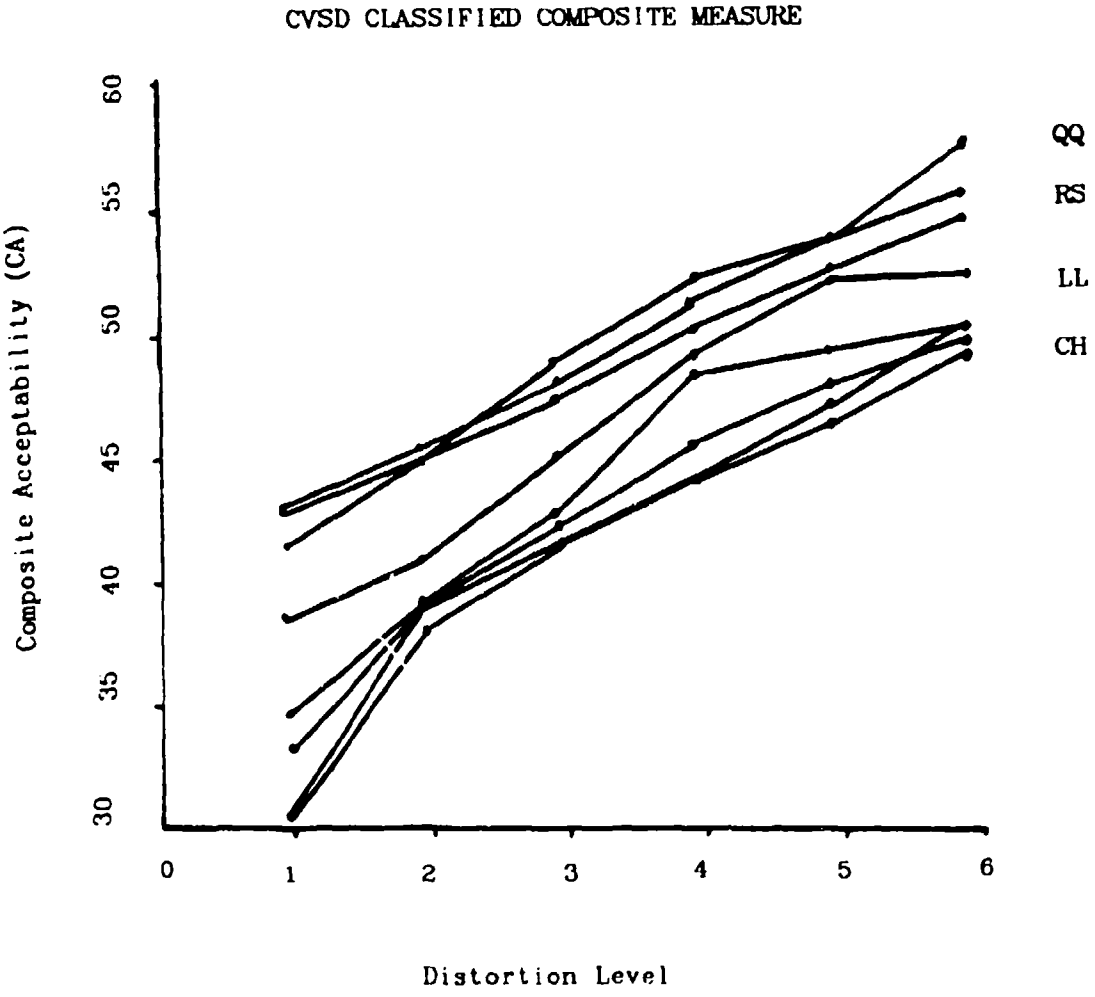


Figure 6.2-7 Composite Acceptability and Estimated Composite Acceptability for CVSD as a Function of Talker and Distortion Level

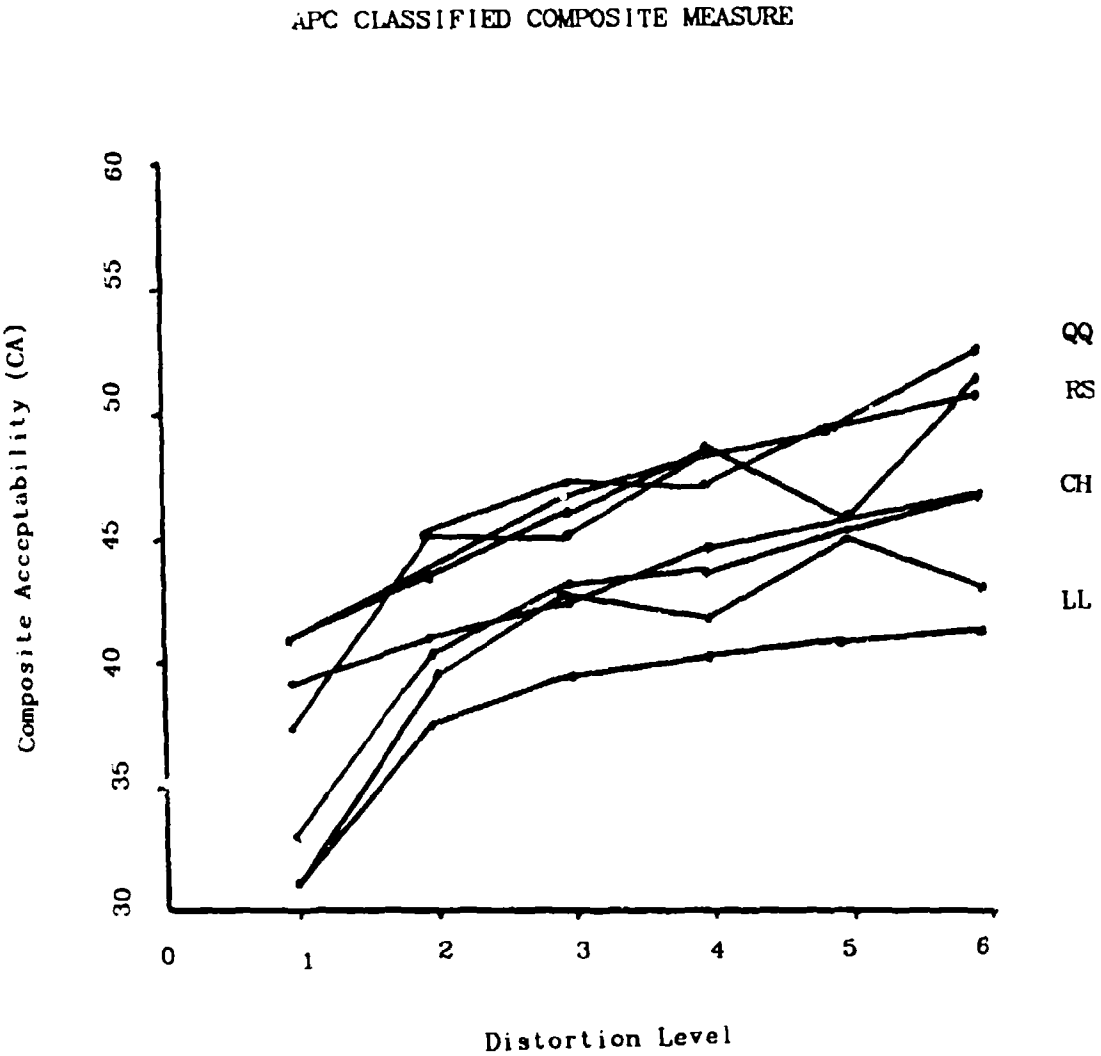


Figure 6.2-8 Composite Acceptability and Estimated Composite Acceptability for APC as a Function of Talker and Distortion Level

distortions. As stated, the two approaches are the same except for the number or type of distortions that are considered in the analysis. The second approach can be simplified, and the two approaches can be made more distinct if the problem is restated as follows: the first approach searches for an objective quality measure given a set of speech distortions, while the second approach searches for a set of speech distortions given an objective quality measure. In both cases the criterion to be satisfied by the search is maximization of the correlation between the objective measure of speech quality and the subjective measure of speech quality over the speech distortions considered. This section reports on work done using the second approach as a means of improving objective speech quality measures.

One can think of the second approach as an objective classification procedure in which speech distortions are objectively categorized into two classes: one class contains the distortions used to assess the objective measure's performance and the other class contains the distortions to be ignored. The approach is similar to that of restricting objective measures to operate only on certain classes of distortions, such as waveform coders; but here the classes of distortions are specified objectively rather than heuristically. The intent is to select a set of distortions objectively which, to a great extent, is homogeneous with respect to the relationship between their objectively measured speech quality and their subjectively measured speech quality. It was hoped that these homogeneous sets of distortions would provide two insights into the objective measure being studied. First, that they would show what kinds of specific distortions are best matched to an objective measure and, second, that they would indicate, by means of common features of the set's members, what overall physical characteristics of the distortions are being measured by the objective quality measure to provide the estimate of subjective speech quality. The next step in this process would be, of course,

to use these insights to adjust or reformulate an objective measure to give a better performance over a given class of speech distortions.

In order to further motivate the approach of searching for homogeneous subsets as a means of improving objective measures, consider an experiment using the log area ratio objective measure. The experiment consists of three regression analyses. In the first analysis a sixth order polynomial regression model was used:

$$CA_i = \beta_0 + \sum_{j=1}^{6} \beta_j O_i^j + \epsilon_i \qquad 6.3\text{-}1$$

in which the objective measure variable, O_i was the log area ratio measure, and the dependent variable, CA_i, was composite acceptability. The regression coefficients, β_j were estimated using the entire set of 44 speech distortions. Subscript j is an index over the order of the model term and subscript i is an index over the 1056 speaker-distortion systems in the distorted speech data base. The resulting correlation of subjective composite acceptability to the regression model's estimated composite acceptability was 0.67, so that the log area ratio objective measure was able to account for only 44.4 percent of the variance of composite acceptability. This result is comparable to the performance of several other simple objective measures, though this performance is not sufficient for providing reliable estimates of subjective speech quality. Table 6.3-1 summarizes these results.

The second regression analysis used the same form as equation 6.3-1, except that the data set was restricted: just four waveform coder distortions were included in the analysis, as specified in Table 6.3-2(a). The results of the analysis, shown in Table 6.3-2(b), are that over the distortion subset specified the log area ratio objective measure was able to account for 49.9

Degree	Regression Coefficient
0	67.21
1	-14.10
2	5.99
3	-1.44
4	.18
5	-.01
6	.00

Multiple R-square .44395

Table 6.3-1 The results using a sixth order polynomial regression model to estimate composite acceptability. The objective measure was the log area ratio distance measure.

Waveform distortions included in the analysis:

Adaptive differential pulse code modulation (ADPCM)
Adaptive pulse code modulation (APCM)
Continuously variable slope delta modulation (CVSD)
Adaptive predictive coder (APC)

(a)

Degree	Regression Coefficient
0	75.73
1	68.94
2	-110.80
3	56.09
4	-12.98
5	1.41
6	-0.06

Multiple R-square .49913

(b)

Table 6.3-2 Part (a) lists the four distortions over which the sixth order polynomial regression analysis was done. Part (b) lists the results of the regression analysis. The objective measure uses was the log area ratio measure.

percent of the variance of composite acceptability. This is a surprisingly small improvement as compared to its performance over the entire set of speech distortions.

The central issue in this experiment is to find out why the log area ratio objective measure performed so poorly over an apparently homogeneous set of waveform coder distortions. One method of investigating this issue is to hypothesize that each distortion conforms to a distinctly different regression model as opposed to a single model as in equation 6.3-1. A means to explore this hypothesis is to use an indicator variable regression model, stated as follows:

$$CA_i = (\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3) + (\beta_4 + \beta_5 Z_1 + \beta_6 Z_2 + \beta_7 Z_3) O_i + \epsilon_i \qquad 6.3\text{-}2$$

Note that this is a linear regression model as opposed to the polynomial regression model used in the previous analysis. The variables Z_j, which have the value either zero or one, are indicator variables, so called because they indicate to which distortion data O_i belongs to as follows:

Z_1	Z_2	Z_3	Waveform Coder Distortion
0	0	0	ADPCM
1	0	0	APCM
0	1	0	CVSD
0	0	1	APC

The indicator variables permit each distortion to have a unique slope and intercept in the regression model. The results of the analysis are shown in Table 6.3-3. The model has improved dramatically, in that it now accounts for 83 percent of the variance of composite acceptability. Hence the hypothesis that each distortion has a unique model was proven true. In particular, Table 6.3-3 shows that coefficients β_5 through β_7 are not statistically different from zero, so that the major difference between models for each distortion is

Variable	Regression Coefficient
0	60.10
1	-4.75
2	-0.23
3	16.10
4	11.29
5	0.94
6	-0.45
7	0.26

Multiple R	.9126
Multiple R-square	.8329

Table 6.1-3 Results of the indicator variable regression model analysis. Again, the objective measure used was the log area ratio measure.

that they each have a different intercept value. This is dramatically illustrated in Figure 6.3-1. The solid lines are the regression curves for each of the four speech distortions. One can see that they have a similar slope but distinctly different intercepts. The dashed curve is the regression curve obtained from the previous sixth order polynomial regression analysis of this data set. The polynomial curve did not represent the underlying model of any of the distortions very well, and hence had poor performance.

What this experiment clearly illustrates is that a heuristically chosen group of speech distortions, such as a group of waveform coders, does not guarantee a homogeneous set of distortions relative to their regression models. It therefore seems reasonable to use a blind statistical approach, as will be discussed in the following section, to select speech distortions which do have similar regression models and can therefore be grouped together and operated on by a given objective measure to estimate subjective composite acceptability.

6.3.2 The Objective Classification Procedure

The distortion classification procedure assumes that the objective measure is specified, and that it is a measure with only one objective measure variable. The objective measures that were considered are a group of the best simple objective measures proposed by Barnwell and Voiers [6.1]. Given the objective measure, the procedure finds the 44 distortion subsets, with number of members one through 44 respectively, which provide the best correlation between the objective measure and subjective composite acceptability. The procedure can be divided into two sections. The first section of the procedure searches through all possible distortion subsets for the subset of size N which provides the greatest correlation between the selected objective measure and composite acceptability. The correlation is computed only over the members of the subset. Let this subset of size N be called S_N. This would be the only section of the procedure were it not for the very large number of computations

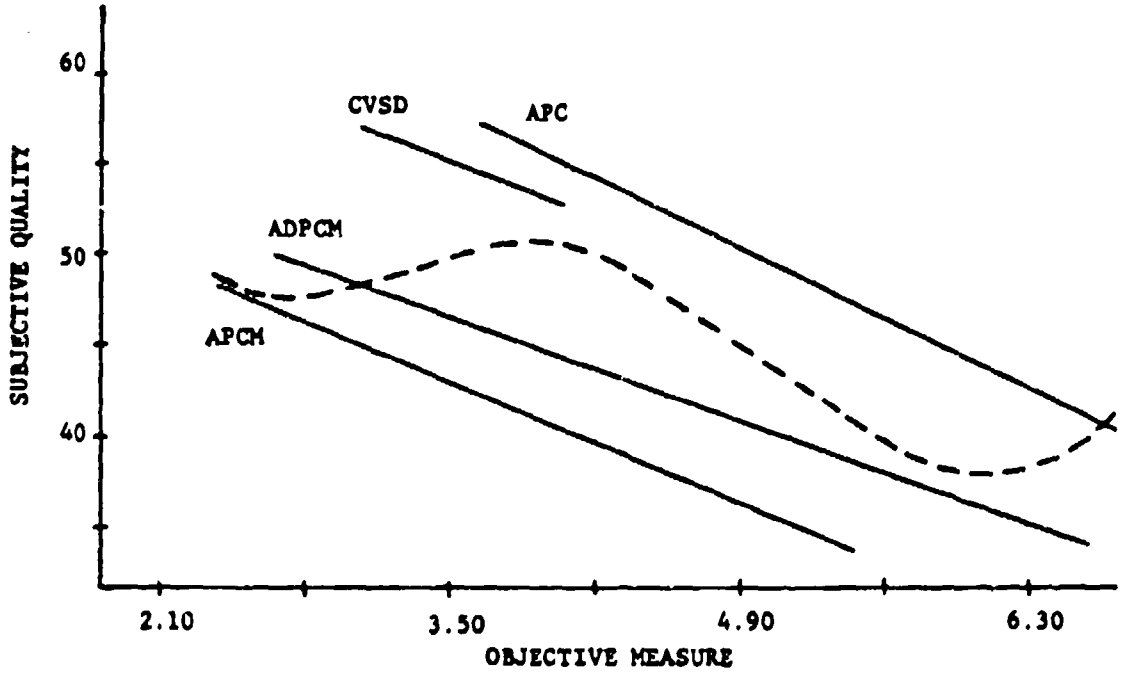


Figure 6.3-1 Each of the four solid curves represents the best linear regression curve fit for each of four distortions. The dashed line represents the best sixth order regression curve fit for all four distortions taken together.

involved as the number of members in the subset grows larger. In the investigation of all subsets of size N the number of correlations that must be computed is equal to the number of combinations of 44 items taken N at a time, or:

$$C_N^{44} = \frac{44!}{(44-N)!\ N!} \qquad (6.3\text{-}3)$$

An exhaustive search of all subsets of all sizes would then require a number of correlation calculations equal to the sum of 44 items taken N at a time for N equals one to 44, a number which exceeds 10^{12}. Because of this excessive number of calculations, the first part of the procedure was only done for subsets of size one through five.

The second part of the procedure circumvents the problem of burdensome calculations at the expense of being sub-optimal. This part searches for a distortion not already a member of set S_N-1 which, when added to S_{N-1}, produces a new set S_N which provides the greatest correlation between the objective measure and composite acceptability. Again, the correlation is computed over the set S_N. This step is repeated for N equals 6 through 44. The entire algorithm is summarized in Figure 6.3.2-1.

6.3.3 Results of Objective Classification into Homogeneous Subsets

The results of the subset classification experiment are, in general, inconclusive. The graph in Figure 6.3.3-1 shows how the correlation coefficient for the best subset varies with the number of members in each subset for each of the objective measures studied. These results look quite promising: for each of the four measures, a subset of fifteen distortions , or one-third of the total number of distortions, had a correlation of better than 0.90. Therefore all of these objective measures are producing very good estimates of subjective composite acceptability for each of the distortions in the subsets. These results are less encouraging when one examines the types of distortions

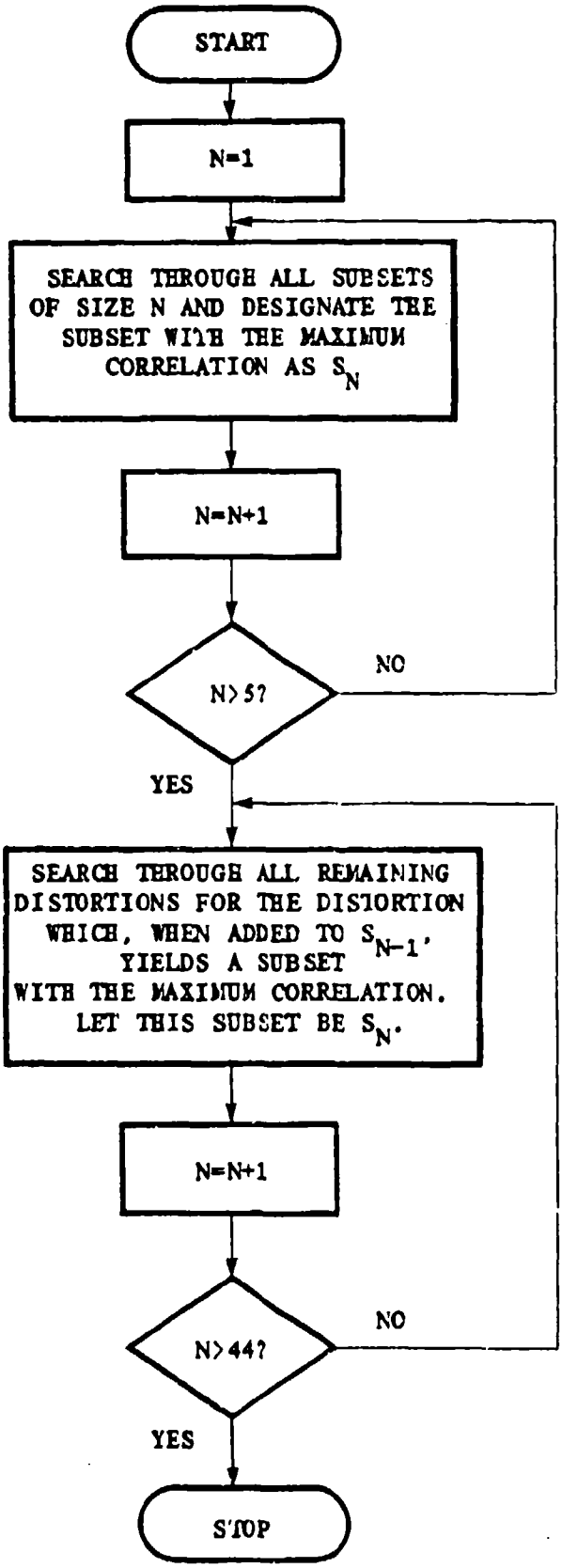


Figure 6.3.2-1 A flowchart illustrating the algorithm used in selecting the best distortion subsets for a given objective measure.

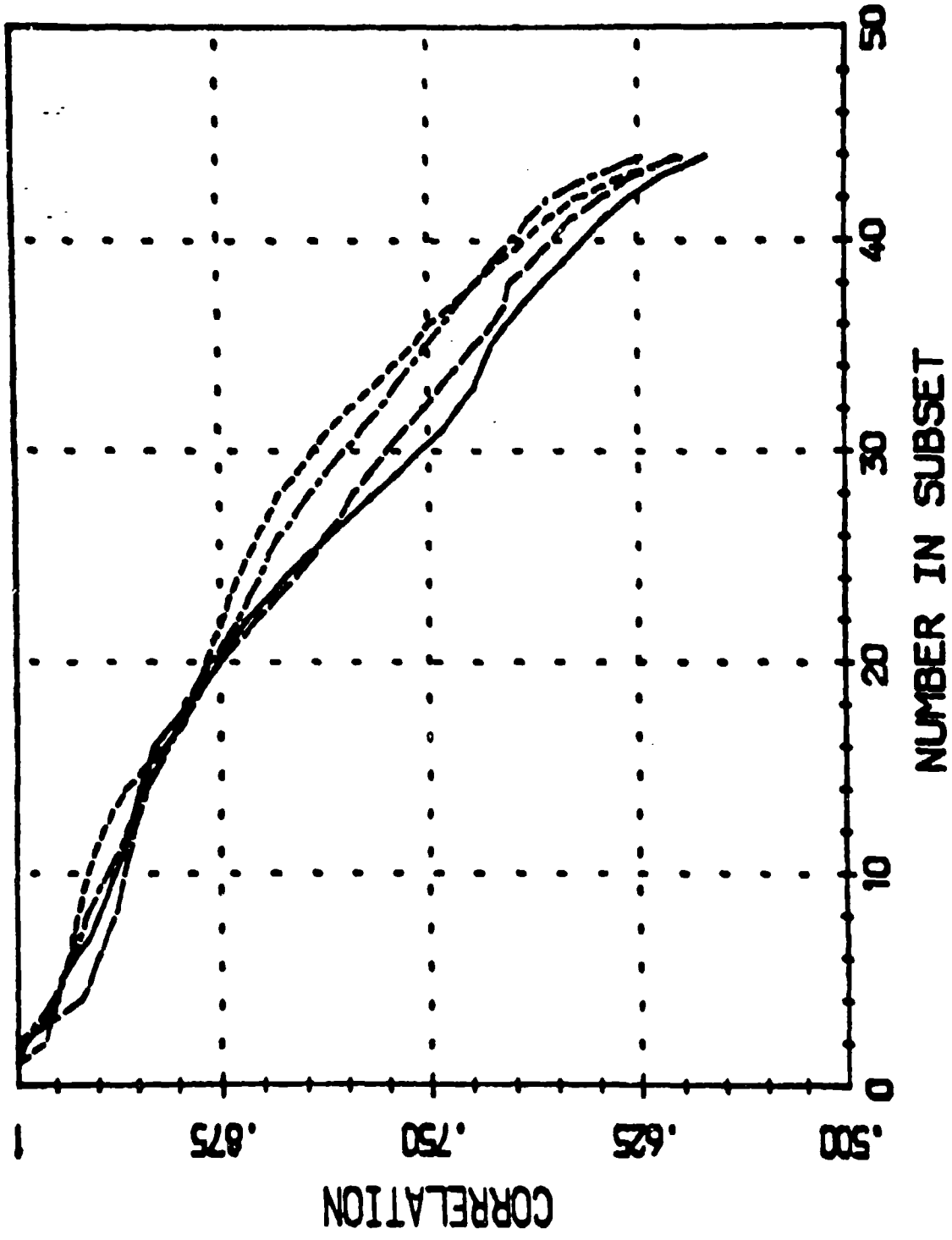


Figure 6.3.3-1 Results of homogeneous subset analysis.

contained in the subsets. Table 6.3.3-1 lists the distortions contained in the subset of fifteen distortions for each of the objective measures presented in Figure 6.3.3-1.

The most remarkable aspect of these subsets is that each contains a very diverse group of distortions. This is quite the contrary of what was hoped in this experiment. A close examination of each subset reveals that there are one or two groups of the same distortion type within each subset. For example, the subset associated with the log spectral distance measure contains three contiguous bands of additive narrowband noise distortions and two contiguous bands of angular pole distortions. Similarly, the subset associated with the Itakura distance measure contains three contiguous bands of additive narrowband noise and four bands of angular pole distortions. The subset associated with the log area ratio distance measure contains three bandlimiting filtering distortions, three contiguous additive narrowband noise distortions and three contiguous banded in-phase noise distortions. Though there are these limited similarities between distortions in the subsets, in general there is not enough commonality between distortions to make any firm conclusions regarding the type of distortions which are best suited for the objective measures. Since it is not clear what general qualities these distortions have in common, it is even less clear what physical qualities of those distortions are being measured to yield the undeniably good estimates of subjective composite acceptability. Hence we are, unfortunately, unable to make hypotheses about the underlying mechanisms which, in a statistical sense, make this set homogeneous.

6.3.4 Conclusions

Intuitively the blind statistical method for choosing homogeneous distortion subsets, as presented in this section, has merit in that it identifies, by the very nature of the algorithm, near-optimal subsets. For all objective measures investigated the performance over subsets containing one-

Log Spectral Distance Measure:	Nonlinear Spectral Distance Measure:
center clipping	800 - 1300 Hz noise
400 - 800 Hz noise	PD 2600 - 3400, frequency
PD 1900 - 2600, frequency	PD 200 - 400, frequency
PD 2600 - 3400, radial	400 - 800 Hz noise
ADPCM	VEV 13
PD 200 - 400, frequency	VEV 7
BD 1300 - 1900	APC
APCM	BD 1300 - 1900
VEV 7	APCM
VEV 13	PD 2600 - 3400, radial
800 - 1300 Hz noise	0 - 400 Hz noise
peak clipping	BD 800 - 1300
PD 1300 - 1900, frequency	center clipping
0 - 400 Hz noise	PD 1300 - 1900, frequency
APC	quantization

Log Area Ratio Distance Measure:	Itakura Distance Measure:
bandpass filtering	800 - 1300 Hz noise
2600 - 3400 Hz noise	BD 1300 - 1900
PD 2600 - 3400, frequency	PD 200 - 400, frequency
PD 200 - 400, frequency	ADPCM
BD 1900 - 2600	center clipping
1900 - 2600 Hz noise	PD 2600 - 3400, radial
BD 100 - 400	APCM
1300 - 1900 Hz noise	PD 1900 - 2600, frequency
PD 2600 - 3400, radial	0 - 400 Hz noise
highpass filtering	peak clipping
BD 800 - 1300	PD 1300 - 1900, frequency
lowpass filtering	BD 100 - 3500
BD 1300 - 1900	400 - 800 Hz noise
APC	PD 800 - 1300, radial
PD 1900 - 2600, frequency	PD 400 - 800, frequency

Table 6.3.3-1 The homogeneous subsets of fifteen distortions for four objective measures. The subsets provide maximum correlation between the objective measure and composite acceptability.

third of the total number of distortions was, in fact, very good, with correlation with composite acceptability exceeding 0.90 in all cases. These facts promote the blind statistical approach as opposed to a heuristic approach to choosing distortion subsets. Unfortunetly, whereas a heuristic approach based on grouping common distortion types, by its very nature, yields physically homogeneous subsets, the blind statistical approach yields subsets which are fragmented, containing small groups of diverse distortion types. This is largely unsatistfying, in that no broad conclusions can be drawn as to the physical or perceptual nature of distortions which are best matched to the objective measure being investigated.

This is not to say that the statistical approach for grouping distortions is entirely rejected, but merely that it is inconclusive based on an initial set of experiments. The conclusion at this stage is, however, that insight into the underlying mechanisms which cause an objective measure to be a good match to a certain set of distortions, and hence permit the objective measure make good estimates of subjective quality, are best found through other experimental approaches. In particular, it is felt that investigation of objective measures for estimating parametric subjective qualities would yield more insight into these issues.

REFERENCES

[6.1] T.P. Barnwell III, and W.D. Voiers, 'An Analysis of Objective Measures for User Acceptance of Voice Communications Systems,' DCA Report No. DA100-78-C-0003, September, 1979.